基于最优匹配模型的数据库压缩算法*

左利云

(茂名学院 实验教学部，广东 茂名 525000)

摘 要：针对数据库中数据急速膨胀的状况，提出一种新的适用于语义压缩的数据库压缩算法——基于最优匹配的 OPMC 算法。算法将数据表中的属性元组分类并进行最优匹配的筛选为每类选取一个代表元组，将数据集中到最优匹配的聚类中心点上，消除相似的、冗余的数据，从而实现数据的压缩。该算法经仿真实验验证，有效改善了压缩比率，相对其他算法的压缩比率提高 18%。

关键词：最优匹配；OPMC 算法；数据压缩

中图分类号：TP392 **文献标识码**：A **文章编号**：1674-7720(2010)14-0014-03

Database compression algorithm based on the optimal matching model

ZUO Li-Yun

(Experiment and Teaching Center, Maoming College, Maoming 525000, China)

Abstract: The data in the database and the rapid expansion of the situation and recommended the application of a new database on the semantic compression algorithms - OPMC algorithm based on optimal matching. Algorithm in the attribute data table tuple classification and make the best match for each type of filter selected a representative tuple will match the data set to the optimal clustering center, eliminating a similar, redundant data, thereby achieve data compression. The algorithm is verified by simulation results indeed effective in improving the compression ratio to increase 18% than other algorithms.

Key words: optimal matching; OPMC algorithm; data compression

数据库正在急速膨胀成应用系统中巨大的组成部分，吞噬着系统的性能。当单一数据库逐步膨胀为 PB 容量时，要查询到适当的存储内容就会越来越困难。每个数据表的容量正在迅速膨胀，数百万行的数据表正在膨胀为数十亿行的大规模数据表，还需要额外的空间来备份所有这些数据。所以，存储访问将更大规模的数据纳入更小的空间是未来数据库面临的巨大挑战。这就需要有更好、更有效率的压缩算法和更高性能的压缩技术。IDG 集团资深专栏作家 Sean McCown 预测网络业的下一件大事就是新的数据压缩技术[1]。

传统的语法压缩方法不能满足大型数据库的需要，因此近年来人们开始研究如何将语义压缩很好地应用于大型数据库。相关研究较早的有 Fascicles[2]算法，是鉴于数据表中一些元组在某些属性值上具有相似性，对其聚合成簇，得到数据的简约表示；ItCompress[3]算法是根据大型数据表中一些元组在某些属性上取值的相似性，提出的一种有损语义的压缩方法。贝尔实验室在大型数据表的基于模型的语义压缩系统 (A Model-Based Semantic Compression System for Massive Data Tables)中提出了 SPARTAN[4]方法，其主要思想是发掘数据属性间的依赖关系、构建属性间的 CART 决策树预测模型和行向聚合成簇。而最新有关数据压缩方法的研究有以消除数据冗余为主要宗旨的 DHFXSC 算法[4]，它通过构造哈夫曼树获得相应的哈夫曼编码，然后匹配产生的哈夫曼编码，该算法的动态构造哈夫曼树是一个比较复杂的过程，会影响压缩效率；而增量型 SDT 算法[5]仅对连续重复字节的压缩和重复出现的字串的压缩比较有效；基于均方误差约束的 MSA 算法[6]需要非常多的计算从而影响压缩速度。

针对以上方法在灵活性和压缩性能方面的缺陷，本文提出一种新的行压缩算法。

1 基于最优匹配模型的 OPMC(最优匹配聚类)算法

本文的压缩算法是通过聚类的方法将数据聚集到较少的聚类中心点上，消除相似的、冗余的数据，从而实现数据的压缩，数据对象主要是数据库中数据表的行数据。

* 基金项目：广东省科技计划项目(2007B010400042)，茂名市科技计划项目(20091009)，茂名学院基金项目(203492)资助

现有聚类算法有 K-均值聚类算法、基于模糊划分的迭代算法、基于遗传算法的最优统计分析(GAC)等。这些聚类方法一般是根据“距离”的度量来确定数据分类的归属，但如果应用在允许一定误差的语义压缩方面，这种基于“距离”的量度则会影响压缩效果。因此提出一种新的最优匹配模型来解决数据分类的归属问题。

1.1 最优匹配模型

一般的聚类模式可表示为如下数学问题。设 $A=\{a_1,a_2,\cdots a_n\}$ 是要进行聚类的数据集，其中 $a_i=[a_{i1},a_{i2},\cdots a_{im}]^T$ 表示第 i 个样本的 m 个属性值。将 A 划分成 k 个子集 $AB_1,AB_2,\cdots AB_k$，在此有 $\bigcup_{i=1}^{k} AB_i=A$，$AB_i \bigcap AB_j=\phi$，其 $i\neq j$。聚类要解决的问题就是如何获得一个最优 k-划分，给出目标函数 min/max $J(A;\boldsymbol{P})$，其中 $\boldsymbol{P}$ 为聚类中心矩阵。

本文的基于最优匹配的聚类算法(简称 OPMC 算法)也采用这种模式。所谓匹配指给定某属性 r 的允许误差 e，通过聚类分组后，如果聚类中心值与实际值在属性 r 的误差在 e 范围内的，称属性 r 上匹配；满足上述匹配条件的属性个数越多，匹配越好。同时也追求最小化存储代价，它包括模型数据和孤立数据。总体匹配程度越好，则独立数据越少。因此要做的就是找到最优匹配。

实现过程如图 1 所示，由允许误差 e，计算与 a_i 最优匹配的 $\boldsymbol{P}$ 中的元组，变量 v 指最优匹配的 $\boldsymbol{P}$ 中元组的下标，sum 指最优匹配的数目。其中外层循环计算 P 中所有元组与 a_i 的匹配数目，内层循环遍历判断每个属性是否匹配，输出最优匹配的元组下标 v 及其数目 sum。

1.2 基于最优匹配模型的 OPMC 算法

OPMC 算法的目标函数为 max $sumH=\sum h(a_i)$，其中 $i=1,2,\ldots,n$；而 $h(a_i)=\max h(a_i,P_i)$，而 $h(a_i,P_i)=m-\sum_{c=1}^{m} g(|a_{ic}-P_{jc}|-e_c)$，其中 $j=1,2,\cdots,k$，当 $a\geq 0$ 时 $g(a)=1$；$a<0$ 时 $g(a)=0$；其中 $h(a)$ 为 a_i 最优匹配属性数目；$sumH$ 为所有元组最优匹配属性数目之和。

Fascicles 算法是针对数据表中一些元组在某些属性值上具有相似性，对其聚合成簇，实现数据的压缩。It-Compress 算法是根据大型数据表中一些元组在某些属性上取值的相似性，将数据表的元组分类并为每类选取一个代表元组，对每一类的任意元组，均用代表元组表示，除非它与代表元组的误差超出了指定范围。

OPMC 算法部分参考了 ItCompress 算法的思想，进行了适当的处理。如在调整聚类中心值的过程中，求取最频繁出现的属性区间，ItCompres 采用分片小区间滑动窗口机制[3]，而本算法采用了两个指针 s 和 t 扫描数据，时间复杂度为 $O(n)$，n 为样本数目。OPMC 算法由数据表及其属性的允许误差 e，求元组归属向量 M/聚类中心矩阵 $\boldsymbol{P}$ 和孤立数据。如图 2 所示，给定初始聚类中心矩阵 $\boldsymbol{P}$，然后循环计算总的匹配数目 $sumH(P)$，并对数据表中的每个元组 A_n 计算最优匹配的 P 中元组，最后在暂时的聚类分组基础上，调整聚类中心矩阵 $\boldsymbol{P}$ 中每个元组的值。

OPMC 算法在压缩前先通过最优匹配模型来筛选，克服了 DHFXSC 算法构造哈夫曼树和 MSA 算法过多计算的影响，可以提高压缩效率，而且 OPMC 算法尤其适合数据

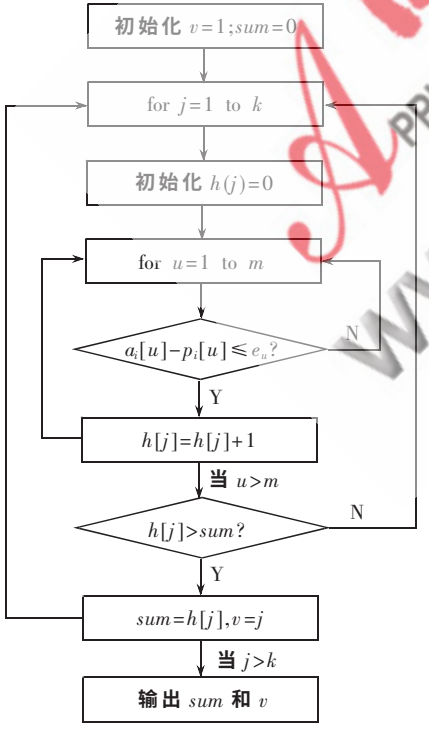

图 1 最优匹配模型实现框图

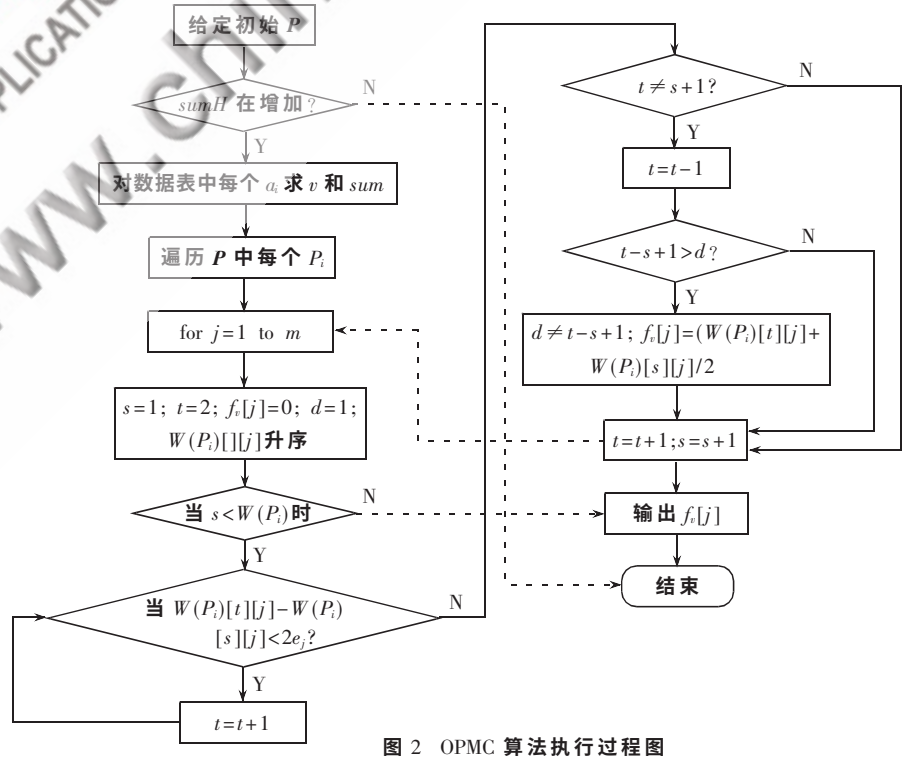

图 2 OPMC 算法执行过程图

属性间线性相关的数据，这点要优于增量型SDT算法。

2 OPMC算法仿真实验

设计了仿真实验来验证算法的性能。因Fascicles算法和ItCompress算法与本算法同是行压缩算法，故将三种算法一起进行实验比较。

实验数据是《华尔街日报》的纽约股票交易所版面上给出的每支股票52周以来每股最高价、最低价、分红率、价格/收益比率、日成交量、日最高价、日最低价、收盘价等信息总共15个属性。实验数据取部分样本数据，样本比率为1%左右，各属性允许误差的设置取值为属性取值范围宽度的一个百分比，如3%。实验首先观察给定允许误差对数据压缩比率的影响，如图3所示。由图可以看出，在数据属性间线性相关关系明显的情况下，采用OPMC算法进行数据压缩，压缩比率比Fascicles和ItCompress算法平均高18%左右(将两算法的压缩比率取平均与OPMC算法相比较)。原因是实验数据中的股票最高价、最低价、分红率等属于连续型数据，Fascicles和ItCompress方法比较擅长处理的是非连续型数据，而且它们只对行进行压缩，而由于数据本身存在明显的属性间线性关系，所以在行压缩前先通过最优匹配模型来筛选，能够提高压缩比率。

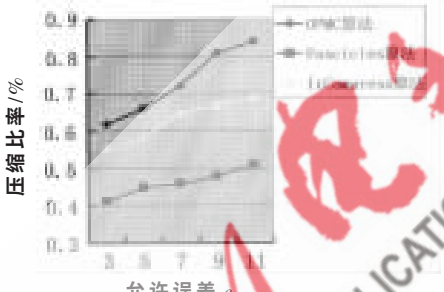

图3 随允许误差变化的压缩比率走势图

图4显示的是OPMC算法的压缩比率随样本比率大小的变化情况，可看出当样本比率大小从0.1%~1.1%变化时，压缩比率变化不大，这说明即使较少的样本数据亦可提取较为理想的压缩模型，由此可见该算法的灵活性。

大型数据库系统急需更好的压缩算法以满足其不断发展的需要，语义压缩比较适合大型数据库系统。本文提出的语义压缩算法——OPMC算法主要适用于连续型数据，特别是存在明显的属性间线性关系的数据。因

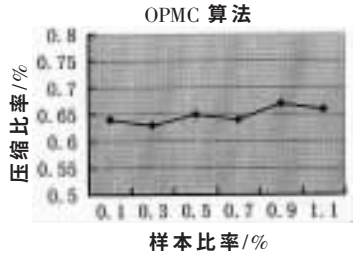

图4 随样本比率变化的压缩比率走势图

其在进行行压缩前先通过最优匹配模型筛选代表元组，从而可以改善提高压缩比率(相对同类行压缩算法Fascicles和ItCompress可提高18%的压缩比率)。本算法不足之处在于对非连续型数据其压缩比率表现并不优于其他三种算法——Fascicles算法、ItCompress算法和SPARTAN方法，对此有待进一步研究改进。

参考文献

[1] MCCOWN S. 网络业未来12件大事[J]. 网络世界，2007(8):11.

[2] VJAGADISH H, MADAR J, NG R. Semantic compression and pattern extraction with fascicles[C]. In Proc. of the 25th Intl. Conf.on Very Large Data Bases. 1999.

[3] VJAGADISH H, RAYMOND TNg, BENG C C,et al. It compress: an iterative-semantic compression mgorithm [C].20th International Canference on Data Engineering (ICDE'04),2004:646-657.

[4] 张晓琳，翟国锋，谭跃生，等. 基于动态哈夫曼编码的XML数据流压缩技术[J]. 内蒙古科技大学学报，2007,26(04):331-336.

[5] 赵利强，于涛，王建林. 基于SQL数据库的过程数据压缩方法[J]. 计算机工程，2008,34(14):58-62.

[6] 高宁波，金宏，王宏安. 历史数据实时压缩方法研究[J]. 计算机工程与应用，2004,40(28):167-170.

[7] BABU S, GAROFALAKIS M, RASTOGI R. Spartan: a model-based semantic compression syatem for massive data tables[C]. In Proc of the ACM SIGMOD'2001 International Conference on Management of Data. May,2001.

(收稿日期：2010-04-17)

作者简介：

左利云，女，1980年生，硕士，讲师，主要研究方向：计算机网络数据库查询及应用。