Available online at www.sciencedirect.com

ScienceDirect

Expert Systems with Applications 36 (2009) 792–803

Expert Systems with Applications

www.elsevier.com/locate/eswa

Estimating null values in relational database systems using automatic clustering and multiple regression techniques

Shu-Ting Chang[a], Shyi-Ming Chen[a,b,*]

[a] Department of Computer Science and Information Engineering, National Taiwan University of Science and Technology, Taipei, Taiwan, ROC
[b] Department of Computer Science and Information Engineering, Jinwen University of Science and Technology, Taipei County, Taiwan, ROC

Abstract

In this paper, we present a new method for estimating null values in relational database systems using automatic clustering and multiple regression techniques. First, we present a new automatic clustering algorithm for clustering numerical data. The proposed automatic clustering algorithm does not need to determine the number of clusters in advance and does not need to sort the data in the database in advance. Then, based on the proposed automatic clustering algorithm and multiple regression techniques, we present a new method to estimate null values in relational database systems. The proposed method estimating null values in relational database systems only needs to process a particular cluster instead of the whole database. It gets a higher average estimation accuracy rate than the existing methods for estimating null values in relational database systems.
© 2007 Published by Elsevier Ltd.

Keywords: Relational database; Null value; Automatic clustering algorithm; Cluster center

1. Introduction

In this information age, many enterprises use relational database systems for storing and processing data. In real-world applications, null values may exist in relational database systems. When some attributes of relational database systems have null values, they will not operate properly. Therefore, how to estimate null values in relational database systems is an important research topic. In recent years, some methods have been proposed to estimate null values in relational database systems (Chang & Chen, 2006; Chen & Chen, 2000; Chen & Hsiao, 2005; Chen & Huang, 2003; Chen & Lee, 2003; Chen & Yeh, 1998; Cheng & Wang, 2006; Huang & Chen, 2002; Lee & Chen, 2002). Chen and Chen (2000) presented a method to estimate null values in the distributed relational databases environment. Chen and Huang (2003) presented a method to generate weighted fuzzy rules for estimating null values in relational database systems using genetic algorithms. Chen and Yeh (1998) presented a method to estimate null values in relational database systems by generating fuzzy rules. Cheng and Wang (2006) presented some methods for estimating null values in relational database systems by utilizing clustering techniques for clustering data and using fuzzy correlations and the distance similarity measure to calculate the correlation of different attributes. Chen and Hsiao (2005) presented an automatic clustering algorithm to estimate null values in relational database systems. Huang and Chen (2002) presented a method for estimating null values in relational database systems with negative dependency relationship between attributes. Lee and Chen (2002) presented a method to estimate null values in relational database systems using genetic algorithms.

In this paper, we present a new method for estimating null values in relational database systems using automatic clustering and multiple regression techniques. First, we present a new automatic clustering algorithm for clustering numerical data. The proposed automatic clustering algorithm does not need to determine the number of clusters

[*] Corresponding author. Address: Department of Computer Science and Information Engineering, National Taiwan University of Science and Technology, Taipei, Taiwan, ROC. Tel.: +886 2 27376417; fax: +886 2 27301081.
E-mail address: smchen@mail.ntust.edu.tw (S.-M. Chen).

0957-4174/$ - see front matter © 2007 Published by Elsevier Ltd.
doi:10.1016/j.eswa.2007.10.046

in advance and does not need to sort the data in the database in advance. Then, based on the proposed automatic clustering algorithm and multiple regression techniques, we present a new method to estimate null values in relational database systems. The proposed method for estimating null values in relational database systems only needs to process a particular cluster instead of the whole database. It gets a higher average estimation accuracy rate than Chen and Chen's method (2000), Chen and Yeh's method (1998), Chen and Hsiao's method (2005) and Cheng and Wang's method (2006) for estimating null values in relational database systems.

2. A new automatic clustering algorithm

Baek and Kim (2007) presented a hierarchical clustering method for learning single-issue negotiation strategies. Saracoglu, Tutuncu, and Allahverdi (2007) presented a fuzzy clustering approach for finding similar documents using a novel similarity measure. Chan, Collins, and Kasabov (2006) presented a greedy K-means algorithm for global gene trajectory clustering. Hsiao and Chen (2001) presented an automatic clustering algorithm to cluster numerical data. However, Hsiao-and-Chen's automatic clustering algorithm only can deal with one-dimensional data and it must sort the data in advance. In this section, we present a new automatic clustering algorithm for clustering numerical data to overcome the drawbacks of Hsiao and Chen's method (2001). The proposed automatic clustering algorithm is now presented as follows:

```
Step 1: Construct the membership functions of linguistic
        terms of the attributes containing null values.
        Define cluster labels based on the values and the
        fuzzified values of the attributes. Cluster the data
        based on the cluster labels preliminarily.
Step 2:/* Choose the tuple which has the smallest total
        Euclidean distance in each cluster as the cluster
        center. */
    for each cluster do
    {
      if the cluster only contains one tuple then the tuple
      is the cluster center of the cluster;
      else if the number of tuples in the cluster is 2 then
      let the mean values of the attributes of these two
      tuples be the cluster center of the cluster
      else
      {
        based on Eq. (1), calculate the total Euclidean
        distance between each tuple and the other tuples
        in the cluster;
        choose the tuple which has the smallest total
        Euclidean distance with respect to the other
        tuples in the cluster as the cluster center of the
        cluster
      }
    }.
Step 3: For each cluster do
    {
      calculate the Euclidean distance between each
      tuple in the cluster and its cluster center,
      respectively;
      calculate the Euclidean distance between each
      tuple in the cluster and the cluster centers of other
      clusters, respectively;
      for each tuple in the cluster do
        {
          if the Euclidean distance between the tuple in
          the cluster and its cluster center is larger than
          the Euclidean distance between the tuple in
          the cluster and the cluster centers of other
          clusters
          then
          {
            remove this tuple from the cluster;
            put this tuple into the cluster which has
            the smallest distance with respect to its
            cluster center
          }
        }
    }.
Step 4: For each cluster do
    {
    if the Euclidean distance between each tuple in the
    cluster and its cluster center is smaller than or equal
    to the Euclidean distance between the tuple in the
    cluster and the cluster centers of other clusters
    then
      go to Step 5
    else
      go to Step 3
    }.
Step 5: Calculate the average distance of each cluster
        based on Eq. (2) and the total average distance
        Total_average_dist of all the clusters based on
        Eq. (3).
Step 6: For each cluster do
    {
      for each tuple in the cluster do
      {
      if the average distance of the cluster > Total_average_dist
      and the Euclidean distance between the
      tuple in the cluster and the cluster center of the
      cluster > Total_average_dist
      then remove this tuple from the cluster
      }
    }.
Step 7: For the removed tuples $data_i$, $data_j$, $\cdots$, $data_p$ do
    {
      if the Euclidean distance between $data_i$ and
      $data_j \leqslant$ Total_average_dist
      then let $data_i$ and $data_j$ form a new cluster,
      respectively
    }.
```

دریافت فوری ← متن کامل مقاله

ISIArticles
مرجع مقالات تخصصی ایران

✓ امکان دانلود نسخه تمام متن مقالات انگلیسی
✓ امکان دانلود نسخه ترجمه شده مقالات
✓ پذیرش سفارش ترجمه تخصصی
✓ امکان جستجو در آرشیو جامعی از صدها موضوع و هزاران مقاله
✓ امکان دانلود رایگان ۲ صفحه اول هر مقاله
✓ امکان پرداخت اینترنتی با کلیه کارت های عضو شتاب
✓ دانلود فوری مقاله پس از پرداخت آنلاین
✓ پشتیبانی کامل خرید با بهره مندی از سیستم هوشمند رهگیری سفارشات