Chapter 7.25
Approximate Processing for Medical Record Linking and Multidatabase Analysis

Qing Zhang
CSIRO ICT Centre, Australia

David Hansen
CSIRO ICT Centre, Australia

ABSTRACT

In this article we investigate how approximate query processing (AQP) can be used in medical multidatabase systems. We identify two areas where this estimation technique will be of use. First, approximate query processing can be used to preprocess medical record linking in the multidatabase. Second, approximate answers can be given for aggregate queries. In the case of multidatabase systems used to link health and health related data sources, preprocessing can be used to find records related to the same patient. This may be the first step in the linking strategy. If the aim is to gather aggregate statistics, then the approximate answers may be enough to provide the required answers. At least they may provide initial answers to encourage further investigation. This estimation may also be used for general query planning and optimization, important in multidatabase systems. In this article we propose two techniques for the estimation. These techniques enable synopses of component local databases to be precalculated and then used for obtaining approximate results for linking records and for aggregate queries. The synopses are constructed with restrictions on the storage space. We report on experiments which show that good approximate

results can be obtained in a much shorter time than performing the exact query.

INTRODUCTION

There is an increasing amount of data being captured in health systems. The data are generally spread among many different data systems. Primarily the data are used for patient treatment; however secondary use of this data is important for improving the health service and medical treatments. These data might be used by the health service to ensure that there is adequate service provision. The medical community may use these data for gathering important information on diagnosis, treatment, and outcome for evidence based practice and future research. With the data spread between so many data repositories, multidatabase systems (Stead, Miller et al., 2000) provide one way to access the data. However, performing queries across large distributed data sets can be expensive, both in time and computing resources.

A typical multidatabase query is to validate patient records across separate local databases which refer to the same person. This can be difficult when, as is often the case, there is no global identifier to provide a single key. For example, suppose one database records the diagnosis information of patients who have been diagnosed with a particular cancer and another database records the information of patients who have received chemotherapy. If no unique patient identifier exists for these two databases, identifying information must be compared from the two databases to identify records which refer to the same patient. Based on the linking results, other analytic queries can then be issued to analyze diagnosis and treatment details.

The linking result can be represented as a *link table* that lists all matched records along with some measure of accuracy of the match. The number of records in the link table is known as the *selectivity* of the link table. The global query optimizer of a multidatabase system uses this selectivity to decide proper strategies on dissecting global query, transferring data among local databases, and so forth. However, a naïve selectivity estimation requires linking two local databases first, which may be very time consuming and even unaffordable due to the large amount of medical data or network failure possibilities. This provides a first motivation for developing an effective selectivity estimation method in a health multidatabase.

Our second motivation comes from the exploratory nature of some queries. In cases, such as decision support systems, users may not be particularly interested in the exact result (Chakrabarti, Garofalakis et al., 2000). An example of this situation may be the case of investigating the relationships between two diseases. Users may first wish to know that the number of patients linked in the two disease databases is big enough to do any further analysis. A large amount of time can be saved if the system can provide an approximate result early. Users then decide if a fuller investigation is warranted.

Motivated by these, we have developed a method for link table selectivity estimation based on constructing database synopses of large and complicate health databases. Since the size of the synopses is strictly limited to the storage requirements of the system, the main challenge in our method is to construct a "good" synopsis which can provide a close estimation to the exact result. We test our approximation methods on various synthetic databases. Experiments show that our techniques can provide close approximation in a much quicker response time than getting the exact result through the naïve query-and-count way.

The rest of the article is organized as follows. The following section presents the background of the selectivity estimation and the structure of our selectivity estimation module. The third section presents our estimation method based on wavelet transformation. The next section presents the