Building small scale models of multi-entity databases by clustering

Georges Hébrail[1] and Yves Lechevallier[2]

[1] ENST Paris, 46, Rue Barrault,
75634 Paris Cedex 13, France, Email: hebrail@enst.fr
[2] INRIA - Rocquencourt, Domaine de Voluceau - Rocquencourt - B. P. 105
78153 Le Chesnay Cedex - France, Email: Yves.Lechevallier@inria.fr

Abstract. A framework is proposed to build small scale models of very large databases describing several entities and their relationships. In a first part, it is shown that the use of sampling is not a good solution when several entities are stored in a database. In the second part, a model is proposed which is based on clustering all entities of the database and storing aggregates on the clusters and on the relationships between the clusters. The last part of the paper discusses the different problems which are opened by this approach. Some solutions are proposed: in particular, the link with symbolic data analysis is established.

1 Introduction and motivation

Every day, more and more data are generated by computers in all fields of activity. Operational databases create and update detailed data for management purposes. Data from operational databases are transferred into data warehouses when they need to be used for decision-aid purposes. In some cases, data are summarized (usually by aggregation processes) when loaded into the data warehouse, but in many cases detailed data are loaded into the data warehouse. This leads to very large amounts of data in data warehouses, especially due to the fact that historical data are kept. On the other hand, many analyzes operated on data warehouses do not need such detailed data: data cubes are often used at a very aggregated level, data mining or data analysis methods only use aggregated data.

The goal of this paper is to discuss methods for reducing the volume of data in data warehouses, preserving the possibility to perform needed analyzes. A important issue in databases and data warehouses is that they describe several entities (populations) which are linked together by relationships. This paper tackles this fundamental aspect of databases and proposes solutions to deal with it.

The paper is organized as follows. Section 2 is devoted to the presentation of related work, both in the fields of databases and statistics. Section 3 describes how several entities and their relationships are stored in databases and data warehouses. In Section 4, it is shown that the use of sampling is not appropriate for two main reasons : (1) the way data must be sampled depends on the goals of further analyzes, (2) the sampling of several populations

and their relationships is difficult to handle. Section 5 presents the model we propose for building small scale models (SSM) of multi-entity databases: the model is based on a clustering of all entities and a storage of information on the clusters instead of the detailed entities. Section 6 discusses the main outcome problems to this approach: (1) the choice of the clustering approach, (2) the updatability of the SSM, (3) the querying of the SSM. Section 7 finally establishes a link between this work and the approach of symbolic data analysis proposed by (Diday (1988)).

2 Related work

This work is related both to the database and statistical data analysis fields.

For many years, work have been done in the field of databases to improve response time for long queries in large databases, or to provide quickly approximate answers to long queries. Summaries of data are built and updated to do so. Two main approaches have been developed: the construction of histograms (see for instance Chaudhuri (1998), Gibbons et al. (1997), Poosala and Ganti (1999)), and the use of sampling techniques (see for instance Gibbons and Matias (1998), Chaudhuri *et al.* (1999), Chaudhuri *et al.* (2001)). In these approaches summaries are built for each single table, not taking into account that tables in relational databases may store either entities or relationships between entities. Our work encompasses the management of summaries of multi-entity databases.

Still in the area of databases, work have been done on the use of data compression techniques in order to improve response time, by storing compressed data on disk instead of original data (see for instance Ng and Ravishankar (1995), Westmann *et al.* (2000)). In this situation compressed data has no interpretation and cannot be used unless decompressing them. Our work differs from this work in the sense that our compression technique has a semantic basis.

At the edge between databases and statistics, much work have been done in the field of (scientific) and statistical databases. In this field, the concept of summary tables (storing aggregated data) has been introduced for many years and some models and query languages have been proposed and developed (see for instance Shoshani (1982), Ozsoyoglu and Ozsoyoglu (1985)). More recently, the concept of multi-dimensional databases has been introduced which enables the user to query interactively summary tables (see for instance Chaudhuri and Dayal (1997)).

In the field of statistics, our work is related to work in two domains: (1) clustering methods, (2) sampling methods. These two domains have been studied extensively for many years (see for instance Duda *et al.* (2001) and Cochran (1997)). Our work and discussion benefits from known results from these domains.

In the field of data analysis, a new direction has been explored at the edge of statistical data analysis and artificial intelligence. The concept of symbolic objects has been introduced to describe objects which are not individuals but have the ability to represent characteristics of groups of individuals. For a complete presentation of this approach, see Bock and Diday (2000). Section 7 shows that a possible solution to describe clusters of entities is the use of the concept of symbolic objects.

Finally, in Hou (1999), it is shown that most of data mining tasks can be achieved using only some summary tables. This work differs from ours in the sense that the summary tables store clusters of individuals which are not based on a clustering process. This approach does not either support multi-entity information.

2.1 Storing several entities in the same database

Relational databases are used in business information systems to store all data needed for the management of the company. For instance data about orders, bills, suppliers, customers, So such business databases store data corresponding to several populations (so-called in the statistical community) known as several entities in the database community.

The information systems community has developed models for describing the contents of such databases. A very famous model, proposed by Chen in 1976 (see Chen (1976)) describes the contents of a database as a set of entities and relationships between entities. More recent models, such as UML (see Booch *et al.* (1999)), are more complete, taking into account the object oriented approach. In this paper, we will refer to a simplified entity-relationship (ER) model proposed by Chen, in order to simplify the presentation. Further work would be necessary to extend to UML, for instance to take into account inheritance.

Within the ER model, the contents of a database are described by sets of entities (for instance cars, customers, products,... as in the example of Figure 1), and relationships between entities. Relationships are represented in Figure 1 by ellipses linked to the corresponding sets of entities. Both entities and relationships can be described by attributes (for instance NAME, ADDRESS for CUSTOMER, and QUANTITY for the relationship linking CUSTOMER, PRODUCT and SUPPLIER). Relationships may represent a relation of arity 2 or more, but in practice no relationships of arity more than 4 appear. Relationships are also described by some cardinality information: in the example of Figure 1, a car is owned by 1 and only one 1 customer and a customer may have from 0 to several cars, a customer may be involved into 0 or several relationships with a couple of products and suppliers. Table 1 describes the attributes of the entities of Figure 1. Note that attributes identifying the entities (called keys) are underlined.

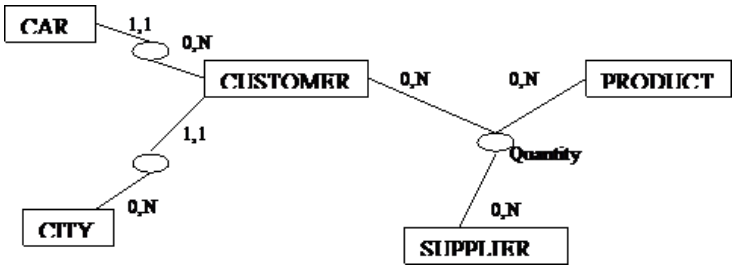

Fig. 1. Example of an ER diagram describing the contents of a database

SETS OF ENTITIES	ATTRIBUTES
CUSTOMER	IDCUST, NAME, ADDRESS, AGE, SALARY, SPC
CITY	IDCITY, NAME, RURAL/URBAN, #POPULATION, #BUSLINES
CAR	IDCAR, NAME, BRAND, PRICE, SPEED, WEIGHT
SUPPLIER	IDSUPP, NAME, ACTIVITY, TURNOVER, EMPLOYEES
PRODUCT	IDPROD, NAME, CATEGORY, SUBCATEGORY, PRICE, WEIGHT

Table 1. Attributes describing entities
Note: SPC stands for Socio-Professional Category

Some rules exist to transform such an ER diagram to a relational database schemata. The diagram above leads to the relational schemata of Figure 2. Each line in Figure 2 corresponds to a table in the relational database. Underlined attributes are primary keys.

CUSTOMER (IDCUST, NAME, ADDRESS, AGE, SALARY, SPC, IDCITY)
IDCITY refers to CITY
CITY (IDCITY, NAME, RURAL URBAN, POPULATION, =BUSLINES)
CAR (IDCAR, NAME, BRAND, PRICE, WEIGHT, IDCUST)
IDCUST refers to CUSTOMER
SUPPLIER (IDSUPP, NAME, ACTIVITY, TURNOVER, =EMPLOYEES)
PRODUCT (IDPROD, NAME, CATEGORY, SUBCATEGORY, PRICE, WEIGHT)
ORDER (IDCUST, IDPROD, IDSUPP, QUANTITY)
IDCUST refers to CUSTOMER
IDPROD refers to PRODUCT
IDSUPP refers to SUPPLIER

Fig. 2. Relational database schemata of the ER diagram of Figure 1

3 Aggregation versus sampling

Considering the example described above, we address now the following problem:

Assuming that the size of the database is very large (for instance millions of customers, cars, products, orders, and thousands of cities), build a small

scale model (SSM) of the database, so that further analyzes may performed on the SSM (computation of cross tables, principal component analysis, clustering, construction of decision trees,...).

For a statistician, a natural way to achieve this goal is to sample the database. Many methods have been developed to do so (see Cochran (1977) for instance): characteristics of a whole population (for instance the average salary of customers) can be approximated by only considering a sample of it. Several problems appear when applying sampling theory to solve our problem. They are discussed below.

3.1 Representativity of the sample

Inference from a sample requires that the sample is large enough and representative. This can be achieved by using simple random or stratified random sampling. Randomization ensures that the inference is correct. Stratified sampling is used when some subsets of the population are too small to be sufficiently represented in the sample, or when data collection makes it necessary. In this case, calibration is performed in order to correct bias. But stratified sampling needs the definition of a goal for the analysis: this constraint is not compatible with our problem since we wish to build a small scale model of the database which can be used for any further analysis. In particular, individuals corresponding to a niche (small sub-population) may be either absent from the sample or not enough represented to consider inference on them.

3.2 Update of the sample

The contents of databases and data warehouses are subject to changes over time. So there is a need for updating the SSM of the database when the initial database changes. Updating a sample is a difficult problem, especially if we want both to keep the sample representative and to limit its size.

3.3 Sampling of several entities and their relationships

The most important difficulty we meet by using sampling as a solution to our problem refers to the management of a SSM describing several sets of entities and their relationships. Since sampling theory applies to one population of individuals, one can sample separately all sets of entities stored in the database. But doing this, there is no chance the relationships between entities are sampled correctly: cars in the car sample do not correspond in general to owners belonging to the customer sample. One can imagine procedures to complement the samples so that the relationships between objects in the sample are complete. But, depending on the cardinalities of the relationships, these procedures may lead to keep the whole database as the sample.

Though efficient algorithms exist to sample relational databases (see Olken (1993)), we consider that this approach is not practical to build SSMs of databases.

4 Small scale model of multi-entity databases

We propose to build the small scale model (SSM) of the database as follows:

- Each set of entities (in our example: CAR, CUSTOMER, CITY,) is partitioned into a large number of classes (typically 10 000 to 100 000) using a clustering algorithm.
- Classes of entities are stored in the SSM, being described by the size the class and by aggregated information associated with entity attributes.
- Relationships between entities are aggregated at the entity class level. They are described by the size of every non-empty combination of classes, and by aggregated information associated with relationship attributes.

Recalling the example, the structure of the SSM, expressed as a relational database, is shown in Figure 3.

SSM_CUSTOMER (C_CUST, D_AGE, D_SALARY, D_SPC, COUNT)
SSM_CITY (C_CITY, D_RURAL URBAN, D_POPULATION, D_=BUSLINES, COUNT)
SSM_CAR (C_CAR, D_BRAND, D_PRICE, D_WEIGHT, COUNT)
SSM_SUPPLIER (C_SUPP, D_ACTIVITY, D_TURNOVER, D_=EMPLOYEES, COUNT)
SSM_PRODUCT (C_PROD, D_CATEGORY, D_SUBCATEGORY,
D_PRICE, D_WEIGHT, COUNT)
SSM_CUST_CITY (C_CUST, C_CITY, COUNT)
C_CUST refers to CUSTOMER
C_CITY refers to CITY
SSM_CAR_CUST (C_CAR, C_CUST, COUNT)
C_CAR refers to CAR
C_CUST refers to CUSTOMER
SSM_ORDER (C_CUST, C_PROD, C_SUPP, D_QUANTITY, COUNT)
C_CUST refers to CUSTOMER
C_PROD refers to PRODUCT
C_SUPP refers to SUPPLIER

Fig. 3. Relational database schemata of the SSM

The interpretation of this relational schemata is the following:

- Each class of entities is identified by a value appearing in a column prefixed by C_. For instance C_CUST identifies classes of customers.
- Classes of entities are described by aggregated descriptions of entity attributes. (prefixed by D_). The simplest way to build such aggregated descriptions is to compute the average value for numerical attributes, and frequency distributions for others. More accurate descriptions can be considered, such as standard deviation or histograms, but are not discussed here. Columns prefixed by D_ correspond to complex data types in the relational database. Note that attributes such as NAME have been removed from the class description since different entities have different names. Note also that all tables describing classes of entities include a COUNT attribute featuring the number of entities in the class.
- Relationships between entities are described by the number of associations of all non-empty combination of classes of entities involved in the relationship. For instance, the tuple ("cluster_car_121", "cluster_cust_342",

26) in table SSM_CAR_CUST means that there are 26 instances of cars of cluster 121 linked to customers of cluster 342 in the detailed database. When relationships also show attributes (like QUANTITY in the ORDER relationship), these attributes are aggregated as entity attributes are.

5 Follow up problems

5.1 Choice of the clustering method

Standard clustering methods do not handle truly large data sets . One solution is to clustering that integrates the Kohonen Self Organizing (SOM) with other clustering methods (see Murthag (1995), Hébrail and Debregeas (1998)). Thus, in the first step, the SOM provides a substantial data reduction, whereby a variety of ascending and divisive clustering algorithms become accessible. The SOM method produces many clusters on the map, the properties of this approach are to reduce the variability of each cluster and to detect some outliers clusters. As a second step, statistical modelling or learning approach provide a framework for model-based clustering. The problem can be treated by a statistical model which assumes a mixture of distributions, each distribution representing one neuron and more complex dependencies can be modelled (Ciampi and Lechevallier (2000)) . An another approach is to associated to a neuron an symbolic object and after the symbolic clustering methods can be easily used. When the data is extracted on the data base they are complex and the type of this data is large (numerical, qualitative, multi categorical,...). In this case a homogenization process can be applied before the clustering step an the strategy is not unique. For instance the fields NAME, ADDRESS contain many attributes and it is necessary to regroup these attributes. One solution is to use metadata information or taxonomy structure defined on the set of attributes by the users.

5.2 Using a database small scale mode

We do not develop here a complete language for querying the SSM but give some examples to show typical queries which can be addressed to a SSM.

Computation of statistics on entities First, SSM tables corresponding to entities can be queried as a whole: for instance the following queries can be answered exactly: *Average salary of customers. Total population in all cities. Total number of bus lines in all cities. Total number of cars. Total number of Renault cars. Number of rural cities.*

Selections may applied to SSM tables corresponding to entities, possibly giving approximate answers to some queries, for instance: *Average salary for customers having SPC = "Farmer"*. To answer this query, all customer

classes with a minimum of 1 customer with *SPC = "Farmer"* are selected (the frequency distribution kept for SPC is used to do so). Selected classes may either contain only farmers or farmers with some other SPCs. An interval of values can then be computed for the average salary over selected classes. Since the clustering has been done with a large number of classes, there is a good chance that most selected classes will contain only or mainly farmers.

Computation of statistics using entities and relationships SSM tables corresponding to relationships can be used to answer queries involving relationships between entities, for instance: *Average salary of farmer customers owning a Renault car.* This query can be answered following the steps below:

- Selection in SSM_CUSTOMER of customer classes with frequency of *SPC = "Farmer"* not equal to zero.
- Selection in SSM_CAR of car classes with frequency of BRAND = *"Renault"* not equal to zero.
- Selection in SSM_CAR_CUST of all couples (car class, customer class) where the car class or the customer class have been previously selected.
- Computation of an interval of values for the requested average salary.

Performing data mining on entities For entities showing only numerical attributes, methods like Principal Component Analysis (PCA) can be applied to the SSM representation, considering that all entities belonging to the same class have the same value for the numerical variable. Intra-class inertia is lost, but this losing is minimized by the clustering process. The same approach can be followed to perform clustering on the SSM. Another way of telling this is to use the ability of most methods to analyze individuals associated with a weight.

When entities show categorical variables, we will see in Section 7 that symbolic data analysis can be applied.

5.3 Updating the small scale model of the database

The update of the SSM is straight forward. When the detailed database changes, the SSM can be updated incrementally in the following way:

- For every created/modified/suppressed entity, its class is computed using a function derived from the clustering process (some clustering methods, like divisive ones, directly give such a function). The SSM table associated with the entity is updated: numerical descriptions are updated (actually the SUM must be stored instead of the AVERAGE value), frequency distribution of categorical attributes are updated by modifying to the right frequency value(s).

- SSM tables associated with the relationships in which the updated entity is involved are also updated (both the COUNT column and possible relationship aggregated attributes).

After many updates, the clustering of entities may become inaccurate: the intra-class inertia may be increasing so that approximate answers to queries addressed to the SSM may be not precise enough. In this case, it may be necessary to re-build the SSM, by applying again the clustering of some entities.

6 Link with symbolic data analysis

Symbolic data analysis (see Bock and Diday (2000)) aims at analyzing objects which describe groups of individuals instead of single individuals. Standard methods of data analysis have been extended to analyze such symbolic objects. A software (called SODAS) is available to run these extended methods. The link between our study and symbolic data analysis can be established at three levels:

- Assertions (which are particular symbolic objects) can used to describe SSMs of entities, but not SSMs of relationships. As a matter of fact, assertions can describe intervals of values, histograms and probability distributions.
- The methods available in the SODAS software can be applied to SSMs of entities.
- The introduction of the SSM structure gives new perspectives for symbolic data analysis: the model of symbolic objects could be extended to take into account relationships between symbolic objects.

7 Conclusion and further work

We have presented in this paper a new model for building small scale models (SSMs) of large databases. Both entities and relationships between entities stored in the database are summarized. This is achieved by clustering entities of the database, and storing aggregates about clusters of entities and relationships between clusters of entities. It has been shown that such SSMs can be used to compute the result of aggregate queries on entities and their relationships. It has also been shown that SSMs can be easily updated when the database changes.

Much further work can be considered on SSMs, mainly:

- The study of relevant clustering methods to build SSMs,
- The definition of a language to query SSMs (inspired from those described in Ozsoyoglu and Ozsoyoglu (2000)),

- The evaluation of the precision of answers obtained by querying SSMs instead of the whole database,
- The extension of symbolic object structure to deal with relationships between assertions.

References

BOCK H.H. and DIDAY, E. (eds.) (2000): *Analysis of symbolic data. Exploratory methods for extracting statistical information from complex data.* Data Analysis and Knowledge Organization, Springer Verlag, Heidelberg.

Booch G., Rumbaugh J., Jacobson I., (1999). *Unified modeling language user guide.* Addison-Wesley, Object Technology Series.

Chaudhuri S., (1998). An overview of query optimization in relational systems, In *Proc. of ACM PODS.*

Chaudhuri S., Das G., Narasayya V., (2001). A robust, optimization-based approach for approximate answering of aggregate queries. *Proceedings of ACM SIGMOD 2001*, May 21-24, Santa Barbara, Californie, USA.

Chaudhuri S., Dayal U., (1997). An overview of data warehousing and OLAP technologies, *ACM SIGMOD Record.*

Chaudhuri S., Motwani R., Narasayya V., (1999). On random sampling over joins. *Proceedings of ACM SIGMOD.*

Chen P.P., (1976). The entity-relationship model towards a unified view of data, In *ACM TODS*, Vol.1, N.1.

Ciampi A., Lechevallier Y., (2000), Clustering Large, Multi-level Data Sets: An approach based on Kohonen elf Organizing Maps, In D. A. Zighed *et al.* (eds.) *Principles of Data Mining and Knowledge Discovery*, PKDD-2000.

Cochran W.G., (1977). *Sampling techniques*, 3rd edition, John Wiley & Sons.

Diday, E. (1988): The symbolic approach in clustering and related methods of data analysis: The basic choice. In: H.H. Bock (ed.): *Classification and related methods of data anlysis. Proc. IFCS-87*, North Holland, Amsterdam, 673–684.

Duda R.O., Hart P.E., Stork D.G., (2001). Chapter 10 : Unsupervised learning and clustering, In *Pattern Classification*, Wiley Interscience.

Faloutsos C., Jagadish H.V., Sidiropoulos N.D., (1997). Recovering information from summary data, In *Proceedings of the 23rd VDLB Conference.*

Gibbons P.B., Matias Y., (1998). New Sampling-Based Summary Statistics for Improving Approximate Query Answers. *Proceedings of ACM SIGMOD 1998.*

Gibbons P.B, Matias Y., Poosala V., (1997). Fast Incremental Maintenance of Approximate Histograms, *Proc. 23rd Int. Conf. Very Large Data Bases, VLDB.*

Hébrail G., Debregeas A. (1998). Interactive interpretation of Kohonen maps applied to curves, *Proc. of the 4th Conf. on Knowledge Discovery and Data Mining* , AAAI press, 179–183.

Hou W., (1999). A framework for statistical data mining with summary tables, In *Proceeding of 11th International Conference on Scientific and Statistical Database Management*, Cleveland (Ohio).

Ng W. K., Ravishankar C.V., (1995). Relational database compression using augmented vector quantization, In *Proceedings of the 11th Conference on Data Engineering*, Taiwan.

Murthag, F., (1995) Interpreting the Kohonen self-organizing feature map using contiguity-contrained clustering. *Patterns Recognition Letters*, 16, 399–408.

Olken F. (1993). *Random sampling from databases*, PhD Dissertation, University of California at Berkeley, USA.

Ozsoyoglu G., Ozsoyoglu Z.M., (1985). Statistical database query languages, *IEEE Softare Engineering*, 12, 1071–1081.

Poosala V., Ganti V., (1999). Fast approximate answers to aggregate queries on a data cube, In *11th Intern. Conf. on Scientific and Statistical Database Management*, Cleveland.

Shoshani A., (1982). Statistical databases, characteristics, problems and some solutions, In *Proc. of 1982 Very Large Data Bases, VLDB.*

Westmann T., Kossmann D., Helmer S., Moerkotte G., (2000). The implementation and performance of compressed databases, *SIGMOD Record*, 29(3):55–67.