The methodology of database design in organization management systems

I L Chudinov, V V Osipova, Y V Bobrova

Tomsk Polytechnic University, 30, Lenina ave., Tomsk, 634050, Russia

E-mail: vikosi@tpu.ru

Abstract. The paper describes the unified methodology of database design for management information systems. Designing the conceptual information model for the domain area is the most important and labor-intensive stage in database design. Basing on the proposed integrated approach to design, the conceptual information model, the main principles of developing the relation databases are provided and user's information needs are considered. According to the methodology, the process of designing the conceptual information model includes three basic stages, which are defined in detail. Finally, the article describes the process of performing the results of analyzing user's information needs and the rationale for use of classifiers.

1. Introduction

Management information systems are among the most important components of information technologies (IT), used in a company. They are usually classified by the functions into the following systems: Manufacturing Execution Systems (MES), Human Resource Management (HRM), Enterprise Content Management (ECM), Customer Relationship Management (CRM), etc. [1]. Such systems are used a special structured database and are required for reengineering of the whole enterprise management system, while the integration makes it difficult to use them. These systems are expensive enough and particularly developed for large enterprises: only in this case the expenses on purchasing and supporting are covered.

As a rule, special information systems (IS) are developed for small and medium sized enterprises. The main goal of the IS is to design and extend the structure of databases, based on the integrated (nonredundant) information environment. Such systems are characterized by their continual developing in expansion of the automated processes. These processes can be implemented effectively if there is a definite methodology for a database design.

2. Approaches to Database Design

Based on the three-level data representation of the ANSI/SPARC architecture [2], the integrated database design process is presented at the following stages:

1. Designing the conceptual information model of the domain area.
2. Selecting the database management system (DBMS), that is used for implementing the database.
3. Presenting the conceptual information model into the physical database structure in notation of the specified DBMS.

The second stage is generally executed only in initial database development, while extending the

information base the DBMS is changed rarely, and concurrent use of different DBMS reinforces the requirement to design of the conceptual information model [3]. For realizing the third stage, there are a number of software tools, e.g. Oracle SQL Developer, Embarcadero ER/Studio XE that automate designing the physical database structure for the specified DBMS, depending on the defined conceptual information model [4]. Whereas the first stage of developing the database is rather considerable, there is no entire formalized description how to design the conceptual information model.

In well-known papers of C. Date [5], R. Barker [6], E. Codd [7,8] designing the conceptual information model is defined to declare the technology as a procedure based on integration of user's information needs. Therefore, the first stage of designing the conceptual information model is considered as the main one in developing a database.

According to the method of learning the domain area, there are two alternative approaches to designing the conceptual information model:

1. Decomposition is based on the system analysis of the domain area.
2. Integration is based on the analysis of current and expectative user's information needs.

The approach, based on the system analysis of the domain area, is much complex, on the one hand, since there is an original methodology to apply or a well-known method of the system analysis to adapt for the purposes of designing the conceptual information model. On the other hand, the decomposition method requires the representative of the domain area, who is familiar with the organization and business function of the company, to participate in developing. Furthermore, the task-specific way of designing the conceptual information model determines applying the integration method. User's information needs are sure to change over time, thus the proposed methodology is focused not only on the initial development of the data schema, but also on integrating a further information need with the current data schema.

3. Methodology of database design

The proposed methodology is to conform to the following principles for designing the conceptual information model:

- Sequential approach to designing that is an incremental modeling of the data schema by including expectative information needs and modifying the current descriptions.
- Noncontiguous interaction of participants in the designing process (suppliers of information requirements and analysts) within a two-level organization of developing:
  - the initial formalization of information needs for the specified tasks;
  - the formalized integration of further requirements description with the base data schema.
- Definition of information needs as structures, related to the typical information formats.
- Declaration of data domains and their relationships as a part of capturing knowledge base about the domain area.
- Use of information about data domains to define availability and type of relationships between data schema components.

The methodology of designing has the following stages:

1. Analysis of user's information needs and their representation as a set of initial entities of the domain area.
2. Definition of initial entities and their representation as plain normalized data schemas.
3. Relation of received normalized entities with the base conceptual information model.

3.1. The first stage of database design

At the first stage, the main semantic analysis of the domain area is performed in terms of user's information needs. Current, as well as expectative information requirements can be classified by means of the following typical information formats:

- Documents are revealed by a special survey of departments.
- Users' requests are revealed by interviewing employees, who fill in documents and use information containing in them.
- Files are revealed by maintenance of the current information processing systems.

Representation of resulting information formats of user's requirements as initial entities requires performing the following actions [9]:

- to identify attributes and the order of them in the entity;
- to identify multiple values of the attribute;
- to define an attribute domain;
- to identify secondary attributes;
- to present the result of attributes identification as a table row of description of initial entities.

Along with the analysis of possible attributes' values, data compatibility is defined:

- = , if attributes consist (domains consist $D1 = D21$).
- < , if attributes of one entity subordinate to attributes of another one ($D1 \subset D2$).
- × , if attributes are comparable ($D1 \cap D2 \neq \emptyset$).
- +, if attributes are compatible ($D1 \cup D2$ makes sense as a domain of the domain area).

3.2. The second stage of database design

At the second stage, the initial entities of the domain area are specified by means of:

- compression of attributes in a new entity;
- normalization;
- identification of classifiers.

Normalization and compression of attributes are formally defined and can be realized easily, satisfying the rules of the relational algebra. Identification of classifiers and decision on whether encoding an attribute are required at the stage of designing the conceptual information model, if the volume of the uncoded view is rather higher than the encoded one, presented as:

$$nl_t > nl_k + m(l_t + l_k) \tag{1}$$

where n – a number of tuples in the entity; m – a number of elements in the dictionary of possible values (in the classifier); l_t– a size of the uncoded (text) value; l_k– a size of the code.

If $l_t = 5l_k$, then

$$nl_k > nl_k + 6ml_k\ , \tag{2}$$

$$nm > \frac{6}{4}. \tag{3}$$

That is, attribute values with the dictionary domain are required to be encoded if a number of tuples in the initial entity is half again a number of possible attribute values.

If there are several (S) attributes in the conceptual information model that are defined at the same data domain, then the condition has the following expression:

$$l_t \sum_{i=1}^{S} n_i > l_k \sum_{i=1}^{S} n_i + m(l_t + l_k), \tag{4}$$

n_i– a number of tuples in the entity, that contains an *i*-th attribute.

If an average number of tuples in the entities that contains encoded attributes, is calculated using formula

$$n' = \sum_{i=1}^{S} n_i S^{-1}, \tag{5}$$

then encoding of attributes is defined by

$$l_t S n' > l_k S n' + m(l_t + l_k). \tag{6}$$

If $l_t = 5l_k$, then

$$5l_k S n' > l_k S n' + 5l_k m + l_k m \tag{7}$$

$$2n'S > 3m. \tag{8}$$

That is, when using such attribute as a full name twice, it is worth applying the classifier, on condition that an average number of tuples in the entity, according to (5), that contain encoded attributes more than (¾ * a number of tuples in the classifier).

3.3. The third stage of database design

At the third stage, the received normalized entities are related in the base conceptual information model. The unified algorithm is proposed to relate separate entities and to receive plain entities in the conceptual information model. For this purpose, relationships between these entities are identified, and constituent entities are analyzed further as independent entities according to the procedure, described in [10].

Furthermore, there are inconsiderable (transitive) relationships between entities identified according the following conditions:

- If there is one hierarchy relationship between entity R1 with key K1 and entity R2 with key (K1, K2), and another hierarchy relationship between entity R2 and entity R3 with key (K1, K2, K3), then the hierarchy relationship between entity R1 and entity R3 can be defined as inconsiderable, transitive.
- If there is a join relationship between entity R1 and entity R2, then the relationships of one subordinative entity can be defined as inconsiderable.

4. Conclusion

The proposed methodology of database design can be applied for developing information systems in any sphere, where objects of organizational management are used. Also the methodology is considered to present the conceptual information model formally, based on current and expected user's information needs.

References

[1] Halevi G 2001 *Handbook of Production Management Methods* (Oxford: Butterworth-Heinemann)
[2] ANSI/X3/SPARC 1975 Study Group on Data Base Management Systems *FDT Bulletin of ACM SIGMOD* **7(2)**

[3] Navathe S B 1992 Evolution of Data Modeling for Databases *Communications of the ACM* **35(9)** 112–123
[4] Kim Y G 1995 Comparing Data Modeling Formalisms *Communications of the ACM* **38(6)** 103–115
[5] Date C J 2003 *An Introduction to Database Systems* (Cambridge: Pearson) pp 59–68
[6] Barker R 1990 *Case*Method: Entity Relationship Modelling* (Wokingham: Addison Wesley Professional) pp 62–75
[7] Codd E F 1970 A Relation Model of Data for Large Shared Data Banks *Communications of the ACM* **13(6)** 377–387
[8] Codd E F 1979 Extending the Database Relational Model to Capture More Meaning *ACM Transactions on Database Systems* **4** 397–434
[9] Osipova V V and Seitz J 2011 A Formal Approach to Design a Conceptual Information Model of the Universe of Discourse *Proc. of the Tenth Wuhan International Conference on E-Business* (New Yourk: Alfred University Press) pp 567–569
[10] Osipova V V, Chudinov I L and Seidova A S 2016 Formalized Approach in Relational Database Design *Key Engineering Materials* **685** 930–933