Construction of an Optimal Relational Database Conceptual Schema using Object-Role Modeling Notation

Davit CHIKOVANI *
Gia SURGULADZE **
Medhat MOUSA ***
Nino TOPURIA ****

Abstract

This article presents the process of conceptual designing of problem area on the bases of object - role modeling and optimal model selection rules. In addition, with respect to the formation of equivalent Entity Relationship ER – models and their comparison issues, in order to optimize the structure of the relational database. After analysis, the optimal variant of designing the relational database structure and quantitative assessment were clarified based on the classical theory of normal forms.

Keywords: conceptual model, Object-Role Modeling, Relational database, ER-model, method of Normal Forms

Introduction

There is no doubt that, the quality of a database application depends critically on its design. Information systems are best specified first at the conceptual level, using concepts and language that people can readily understand. It helps us to fix the semantics of the information received from the user and realize the model on different platforms. The user can describe problem area in various ways. The present work is initiated for finding the best variant of data modeling.

Conceptual modeling using Object - Role Modeling (ORM), ORM simplifies the design process by using natural language, as well as intuitive diagrams, that is populated with examples and by examining the information in terms of simple or elementary facts. By expressing the model in terms of natural concepts, like objects and roles, it provides a conceptual approach to modeling (Halpin & Morgan, 2008).

The Problem Statement

Conceptual scheme may change (many times) in the equivalent conceptual scheme in order to get more efficient logic scheme. In other words, the conceptual optimization is considered in the present work in order to choose a good logical scheme that could improve a database application to work with, in particular to accelerate the requirement.

Consider a simple example. Objects of survey are selected from a university department. The necessary elementary facts to construct ORM-diagram are as follows:

f1: Lecturer in the department can be Assist. Prof. Dr.;
f2: Lecturer in the department can be Assoc. Prof. Dr.;
f3: Lecturer in the department can be Full Prof. Dr.;
f4: Subject can be delivering by one or two lecturers;
f5: Subject has credit;
f6: Lecturer has staff salary;
f7: Lecturer has academic salary per hour;
f8: Lecturer has the rank from university;
f9: Lecturer has the rank …. this year.

Subsett ⓒ constraint indicates that if it is written the date of receipt of the lecturer's rank if and only if the lecturer has earned this title.

Exclusive - or ⊗ constraint is simply an orthogonal combination of an inclusive-or (disjunctive mandatory) constraint and an exclusion constraint. By default, these two constraints are overlaid as shown in Figure 1. Exclusive-or constraint shows that the lecturer may have engaged one of these positions (an assistant professor, an associate professor or a full professor).

* Assist. Prof. Dr. Faculty of Computer Technology and Engineering, International Black Sea University, Tbilisi, Georgia.
Email: dchikovani@ibsu.edu.ge; ibsu.geo@gmail.com
**Prof. Dr., Faculty of Informatics and Control Systems, Georgian Technical University, Tbilisi, Georgia.
Email: g.surguladze@gtu.ge
*** Dr, Program Coordinator of Informatics at Computer Technologies and Engineering Faculty, International Black Sea University,
E-mail: mmousa@ibsu.edu.ge
****Assoc. Professor Dr., Faculty of Informatics and Control Systems, Georgian Technical University, Tbilisi, Georgia.
Email: n.topuria@gtu.ge

Fact of ORM above the graph looks like:

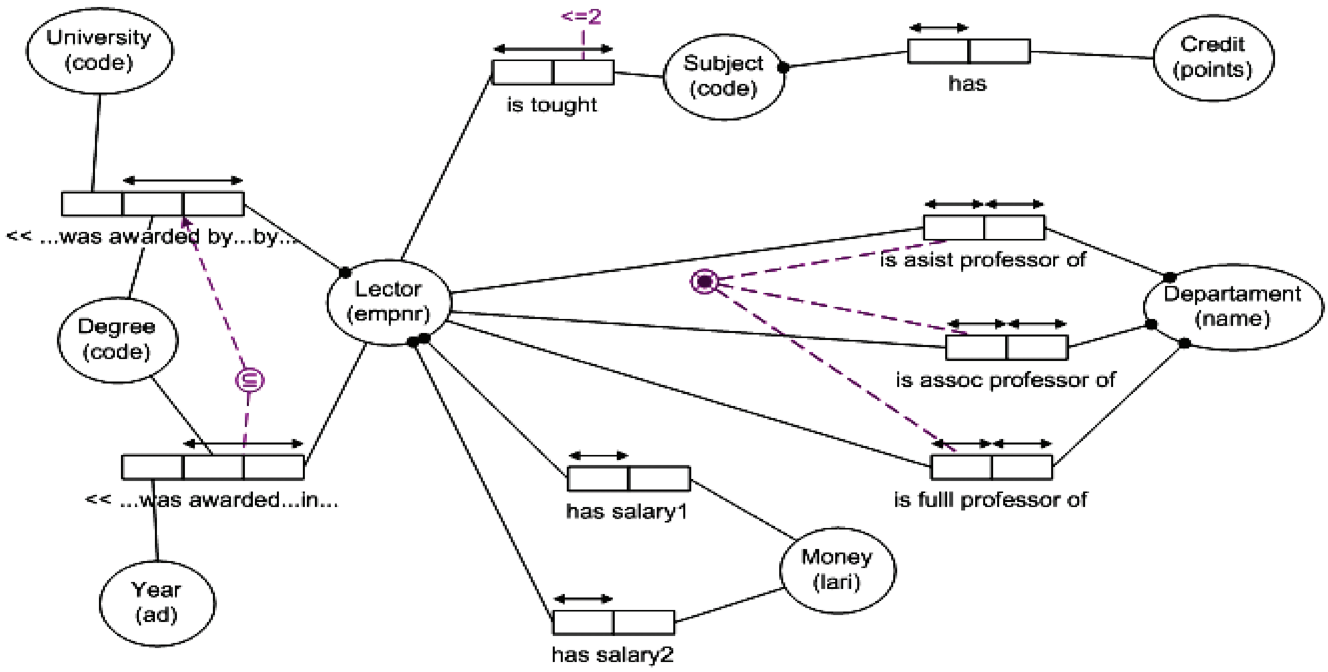

Figure 1. *Fragment of diagram of ORM*

Diagram of the ORM corresponding diagram of ER looks like:

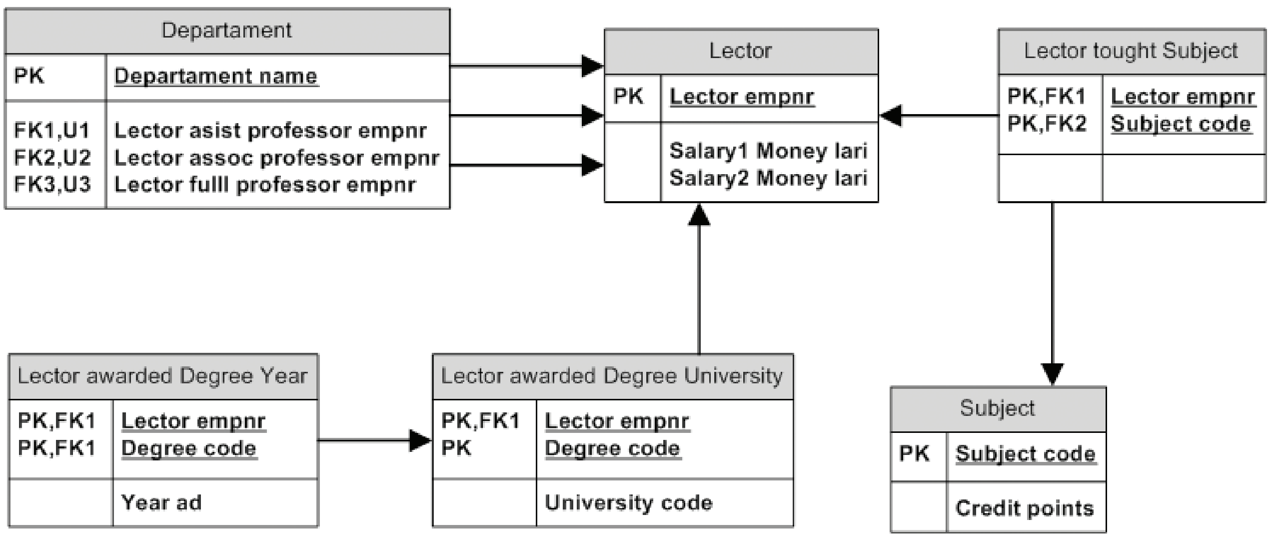

Figure 2. *Fragment of diagram of ER*

Now, a different way to present the facts of the problem area:

f1: Lecturer has the rank (Assist. Prof. Dr.; Assoc. Prof. Dr; Full Prof. Dr.;
f2: Lecturer is working in department.
f3: Lecturer has salary (staff or per hour)
f4: Lecturer has the rank;
f5: Lecturer has the rank from the university;
f6: Lecturer has the rank …. in this year.
f7: Teacher is delivering other subject.
f8: Subject has credit.

Journal of Technical Science and Technologies; ISSN 2298-0032; Volume 3, Issue 2, 2014

An exclusion ⊗ constraint between roles sequences indicates their populations should always be disjoined (mutually exclusive). Figure 3 represents a pair-exclusion constraint - no lecturer is the first and the second lecturer of the same subject.

Subset ⊆ constraint indicates that if the subject has a second lecturer, then that subject also has a first lecturer.

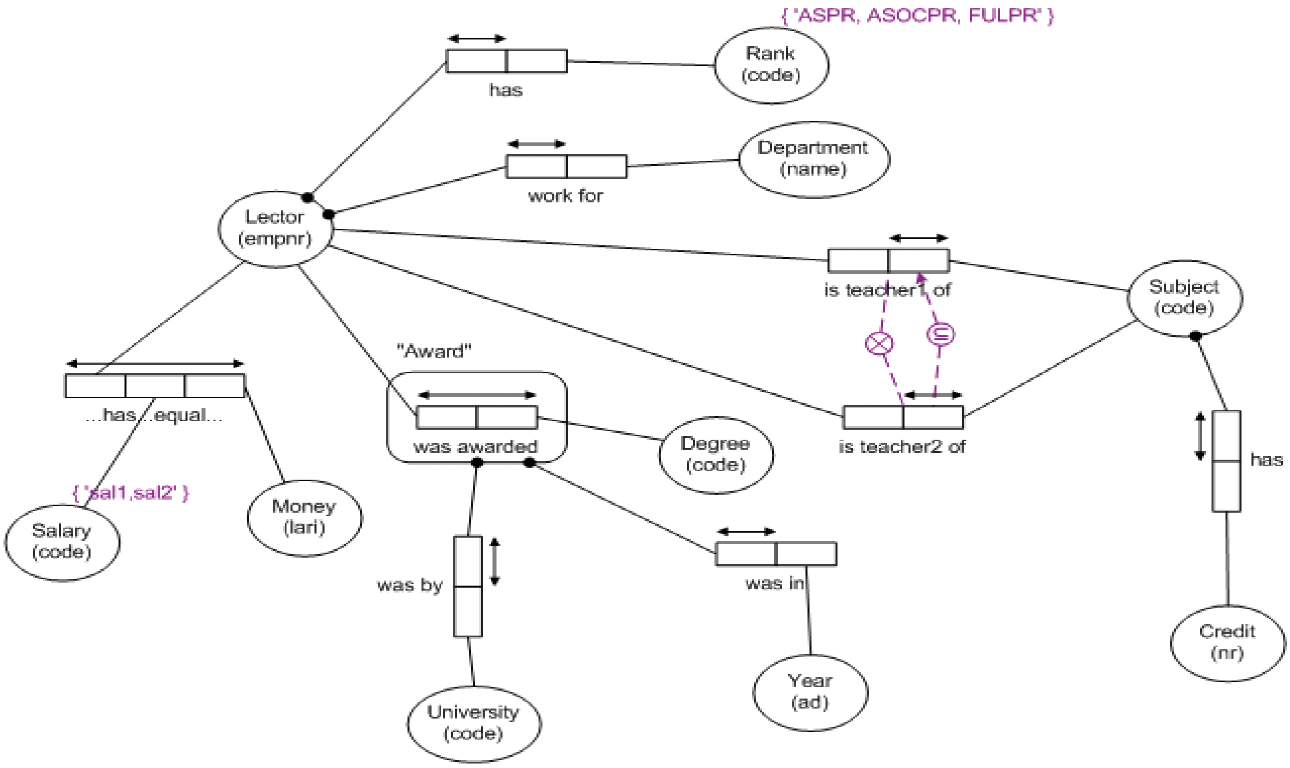

Figure 3. *Fragment of diagram of ORM*

Diagram of the ORM corresponding diagram of ER looks like:

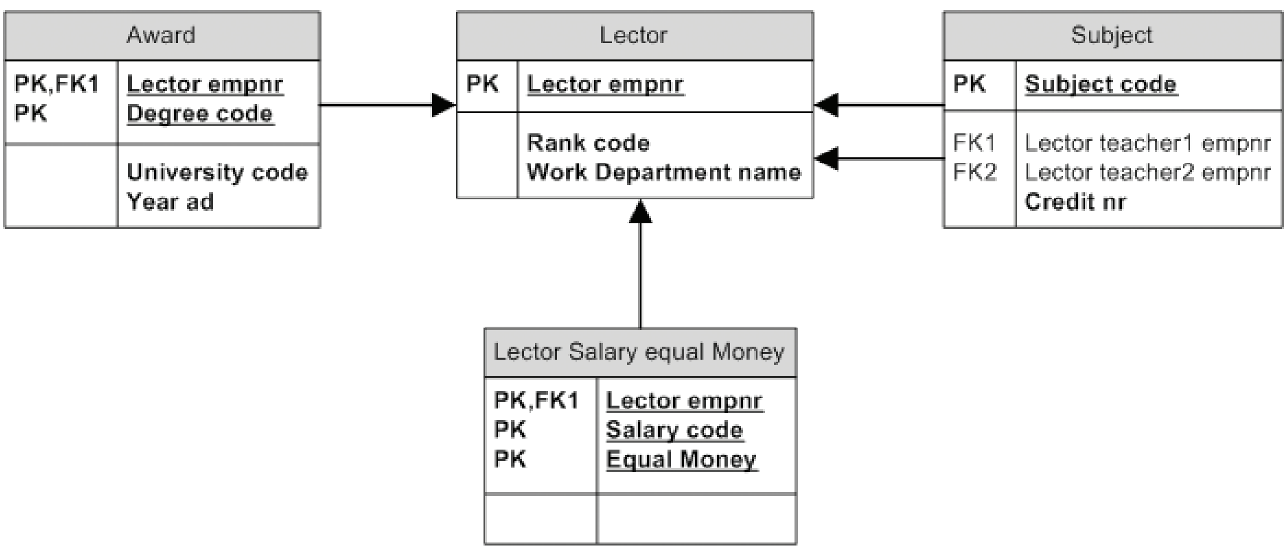

Figure 4. *Fragment of diagram of ER*

The comparison of ERP diagram reveals that change of ORM model has simplified the ER model; in particular, instead of six tables we now have four. That is, optimization entails minimization of quantity of tables during the process of ER model construction, when semantic integrity is sustained. In other words, this is a certain reduction of informational abundance. Here a question may be raised - what quantity of tables and of what structure is optimal? How can we conduct a quantitative assessment taking into account relations, attributes, domains, analysis of their usage and change frequency.

Practice has witnessed that sometimes conceptual changes necessary for optimization can make the model more complex for the designer. Following is also acceptable, that one conceptual scheme might be used for easily constructing the model content, while another - for efficiency of realization.

The problem of determining optimal composition of ER model tables is linked with normalization of structure of relational database (E.F. & Cliffs, 1972), (Fagin., 1981), (G. Surguladze, 1983), (Halpin, & Morgan, 2008), (G. Chogovadze, 1996). Solution to this problem and design of respective algorithm are given.

Methodology

Suppose that $U=\bigcup_i U_i$ is a set of attributes describing the problem area. We can figure the original scheme of database $F^0(\overline{A})$ in the form of universal dependence: $\overline{S}^0=\{\overline{R}< U,P>\}$ where $\overline{R}$ is universe of dependencies, while is classes of semantic restrictions or unity of functional dependencies (FD)(Chogovadze, 2001), (G. Chogovadze G. S.,1996).

The objective is to construct an equivalent to $\overline{S}^0$ scheme $\overline{S}$: $\overline{S}=\{R_i=<U_i,P_i>\}$, where R_i is projection of $\overline{R}$, and i correspondes to a defined class of FD, e.g., FD, Full FD (FFD), transitive functional dependencies (TFD), pseudo-transitive FD (PTFD), Multivalued (MVD), general non-functional (GNF).

Sequential decomposition of the universe is carried out using the theory of Normal Forms (NF) to the following scheme:

NNF -> 1NF -> 2NF -> 3NF -> 4NF -> 5NF -> . ? . -> BNF,

Where NNF is Non-normalized form, 1NF – first N, and BNF is Binary NF.

We may consider Boyce-Codd NF (from 3NF) as a branch of the main scheme, first-degree hierarchical decomposition (from 4NF), interdependencies (from 5NF), etc. The question mark sign „?" indicates that research on depenency attributes is still carried on with the objective of their further optimization.

At present, there are many other types of NFs, but they are rarely reflected in the classic theory of normalization (G. Chogovadze G. G., 2001).

During decomposition of dependencies, BNFs are obtained at each stage of scheme design. For non-decompositional dependencies we need to introduce fictious attributes (a finite set of natural numbers). Semantic integrity among decomposition dependencies is ensured by means of duplicating index links, i.e. certain redundancy is introduced. The higher the NF number, the smaller the informational and the higher the indexical redundancy:

1NF -> 2NF ->. . . -> BNF,
V1_inf >=V2_inf >=...>=VB_inf,

1NF -> 2NF ->. . . ->BNF,
V1_ind <= V2_ind <=...<= VB_ind

Consequently, compromising task was set to determine the optimal redundancy with the requirement to minimize renewal time. Once the problem is solved, as a result of it, it becomes possible to determine the optimal NFs for database scheme dependencies, after which automated design of structures of Normal Forms will be done.

Suppose a set of semantically compatible FDs is given:

$$\begin{cases} R_1(k_1,k_2,\ldots k_{n1},A_1,A_2,\ldots,A_{a1}) \\ R_2(k_1,k_2,\ldots k_{n2},B_1,B_2,\ldots,B_{a2}) \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ R_l(k_1,k_2,\ldots k_{nl},Z_1,Z_2,\ldots,Z_{al}) \end{cases} \quad (1)$$

Where k is a key while A -:-Z is a non-key part.

Semantic integrity is a result of the fact that the set is derived by decomposing a single universum. If we assume that:

$$k_1,k_2,\ldots k_{n1} \supseteq k_1,k_2,\ldots k_{n2} \supseteq k_1,k_2,\ldots k_{nl}$$

(1) by composing the system, it is possible to get comparably lower NF:

$$R(k_1\ldots k_{n1},A_1\ldots A_{a1},B_1\ldots B_{a2},\ldots Z_1\ldots Z_{al}) \quad (2)$$

Also, assume that number of change of R_i which is $\square_i$ is known in advance within defined time interval and following arrangement is true:

$$\mu_1 \geq \mu_2 \geq \ldots \geq \mu_l$$

Expressions (1) and (2) volume of renewals are consequently calculated as follows:

$$Q_{dec}=\sum_{i=1}^{l}\mu_i*(n_i+a_i)$$

and

$$Q_{com}=\mu_1*(n_1+\sum_{j=1}^{l}(a_j-r))$$

where r is such a number of attributes using which join operation is carried out (can be ignored later).

If we assume that there exists an intermediary NF between NFs (1) and (2), then for such, volume of renewals will be:

$$Q=\sum_{j=1}^{s}\mu_j*(n_j+\sum_{k=1}^{l}a_k)$$

Where S is quantity of FDs within the intermediary NF. Following inequality holds true:

$$\mu_1*(n_1+\sum_{k=1}^{l}a_k)\geq\ldots\geq\sum_{j=1}^{s}\mu_j*(n_j+\sum_{k=1}^{l}a_k)\geq\ldots$$

$$\ldots\geq\sum_{i=1}^{s}\mu_i*(n_i+a_i)$$

(3)

Where left edge part of inequality corresponds to (i-1) NF, right edge part corresponds to (i+1)NF, and middle part to the i NF, where *i>=4*.

Let us analyze in detail two adjacent NFs, e.g. i and i+1. Following can be derived from (3):

$$\sum_{j=1}^{s}\mu_jn_j+\sum_{j=1}^{s}\sum_{k=1}^{l}\mu_ja_k\geq\sum_{i=1}^{l}\mu_in_i+\sum_{i=1}^{l}\mu_ia_i$$

(4)

of which following expressions are true:

$$\sum_{i=1}^{l}\mu_in_i-\sum_{j=1}^{s}\mu_jn_j=\sum_{i=s+1}^{l}\mu_in_i$$

(5)

$$\sum_{j=1}^{s}\sum_{k=1}^{l}\mu_ja_k-\sum_{i=1}^{l}\mu_ia_i=\sum_{j=1}^{s}\sum_{k=1,j\neq k}^{l}\mu_ja_k-\sum_{i=s+1}^{l}\mu_ia_i$$

(6)

If we put the right edge parts of (5) and (6) equalities into (4), we get:

$$\sum_{j=1}^{s}\sum_{k=1,j\neq k}^{l}\mu_ja_k\geq\sum_{i=s+1}^{l}\mu_in_i+\sum_{i=s+1}^{l}\mu_ia_i$$

(7)

Let us divide both sides of inequality by $\sum_{i=s+1}^{l}\mu_ia_i$, we get:

$$\frac{\sum_{j=1,k=1}^{s}\sum_{j\neq k}^{l}\mu_ja_k}{\sum_{i=s+1}^{l}\mu_ia_i}\geq\frac{\sum_{i=s+1}^{l}\mu_in_i}{\sum_{i=s+1}^{l}\mu_ia_i}+\frac{\sum_{i=s+1}^{l}\mu_ia_i}{\sum_{i=s+1}^{l}\mu_ia_i}$$

(8)

Since $[1:l]=[1:s]\cup[s+1:l]$ therefore:

$$\frac{\sum_{j=1}^{s}\sum_{k=1,j\neq k}^{l}\mu_ja_k}{\sum_{i=s+1}^{l}\mu_ia_i}=\frac{\sum_{j=1}^{s}\sum_{k=1,j\neq k}^{s}\mu_ja_k}{\sum_{i=s+1}^{l}\mu_ia_i}+\frac{\sum_{j=1}^{s}\mu_j\sum_{k=s+1}^{l}a_k}{\sum_{i=s+1}^{l}\mu_i\sum_{i=s+1}^{l}a_i}$$

Thus, we get Wedekind-Surguladze model from (8) (G. Chogovadze G. G., 2001) (G. Surguladze, 1983):

$$\frac{\sum_{j=1}^{s}\sum_{k=1,j\neq k}^{l}\mu_ja_k}{\sum_{i=s+1}^{l}\mu_ia_i}+\frac{\sum_{j=1}^{s}\mu_j}{\sum_{i=s+1}^{l}\mu_i}\geq\sum_{i=s+1}^{l}\frac{n_i}{a_i}+1$$

(9)

Where $l\geq2,s\geq1$ and $l>s$.

The following case is often used in practice, where l=2 and s=1, then (9) takes the following form:

$$\frac{\mu_1}{\mu_2}\geq\frac{n_2}{a_2}+1$$

(10)

Which, as it is known, is the Wang-Wedekind model (Wang, January, 1975.). It is a particular case of expression (9). Usage range of (10) is up to 3NF, while for (9) it is complete range of NFs, thus, it is universal.

Now let us study a case with high change frequency of non-key attribute values, while key ones – with low frequency. Suppose *l=2, s=1* and schemes of relations are given:

$$\begin{cases}R_1(k_1,k_2,\ldots k_{n1},A_1,A_2,\ldots,A_{a1})\\R_2(k_1,k_2,\ldots k_{n2},B_1,B_2,\ldots,B_{a2})\\ \\R_{12}(k_1,k_2,\ldots k_{n1},A_1,A_2,\ldots,A_{a1},B_1,B_2,\ldots,B_{a2})\end{cases}$$

Where following conditions are true

$$k_1,\dots k_n \supseteq k_1,\dots k_{n2} \text{ and } \mu_1 > \mu_2 \qquad (11)$$

Following (9), for the given schemes of R_1, R_2 and R_{12}, in case of high frequency change of key attribute values R_1 and R_2 recommended within the R_{12} composition of dependencies, if following condition is met:

$$\mu_1(n_1+a_1)+\mu_2(n_2+a_2) > \mu_1(n_1+a_1+a_2)$$

We can deduce that:

$$\frac{n_2}{a_2} > \frac{\mu_1}{\mu_2} - 1$$

If value changes of non-key attributes are considered without changes of key attributes, then following expression is true:

$$\mu_1 a_1 + \mu_2 a_2 > \mu_1(a_1+a_2)$$

We can deduce that:

$$\mu_2 > \mu_1$$

this contradicts (11).

Thus, dependency schemes experiencing dominant non-key attribute part changes are recommended to be expressed using high range NFs.

Conclusion

• Relational database scheme can be effectively constructed using ORM-tool. In order to compare the equivalent schemes and find the optimal scheme it is necessary to use the quantitative evaluation algorithm, that is designed based on the classical theory of Normal Forms of databases;

• The scheme transition on the conceptual level may be used to throw more light upon the conceptual model or improve quality of database application;

• Database dependencies should be presented using NFs of different degrees (3NF-:-BNF), within given context, depending on renewal frequency and link types;

• For the given expression (9) can be used to determine optimal NFs;

• If change frequency of key dependencies attributes is high, then, it is reasonable to use low degree of NFs, and if change frequency of non-key attributes is the dominating one, then NFs of comparably higher degree are to be used.

References:

C. Wang, H. W. (January, 1975). Segment Synthesis in Logocal Data Base Design. IBM J. RSD 19, N1, p.71-77. .

E.F. , C., & Cliffs, E. (1972). Further normalization of the database relational model. In Data Base Systems, Courant Computer Science Symposia 6. Prentice-Hall, 65-98.

Fagin., R. (1981, September). A Normal Form for Relational Databases That Is Based on Domains and Keys. IBM Research Laboratory. ACM Transactions on Database Systems, Vol. 6, No. 3,, pp. 387-415.

G. Chogovadze, G. G. (2001). Design and Building of the Management Information Systems. p 742. Tbilisi, ,: GTU.

G. Chogovadze, G. S. (1996). Bases of Construction of Database and-Knowledge. . Tbilisi State University. “Ganatleba”. Tbilisi, , 375 .

G. Surguladze, V. K. (1983.). On the Choice of Acceptable normal forms for Logical Relational Databases. ransact. of GTU, “Techn. Cybernetics”, N 10(267), Tbilisi., pp. 47-51.

Halpin, , T., & Morgan., T. (2008.). Information Modeling and Relational Databases. Second Edition, The Morgan Kaufmann Series in Data Management Systems, 943 p.