VALIDATING REFERENTIAL INTEGRITY AS A DATABASE QUALITY METRIC

CORAL CALERO, MARIO PIATTINI, MACARIO POLO, FRANCISCO RUIZ
Grupo ALARCOS
Universidad de Castilla-La Mancha. Spain
34 26 295300 ext 3731
e-mail: {ccalero, mpiattin, mpolo, fruiz} @ inf-cr.uclm.es

ABSTRACT:[†]

This paper describes two metrics based on the referential integrity. The first one is defined as the maximum number of levels of referential integrity among tables and the second one is defined as the number of foreign keys. An empirical study for demonstrate that these metrics can affect the understandability of the relational database schema is presented. Four cases have been designed in order to validate empirically the influence of the two metrics. With the results obtained we conclude that the referential integrity affects the undestandability of the relational database schema.

Keywords

Empirical software engineering, referential integrity, database metrics.

1. INTRODUCTION

Database quality depends on several factors, one of which is maintainability [7]. Maintenance is considered the most important concern for modern Information Systems (IS) department and requires greater attention by the software community ([6], [9], [13]). It is neccesary to measure software maintainability in order to reduce the amount of effort expended in software maintenance activities [10].

Maintainability is affected by understandability, modifiability and probability which depend on complexity [8].

In this paper two different metrics, based on the referential integrity, are proposed to measure database product complexity. In section 2, we describe the two metrics. Metrics verification in the [3] framework is presented in section 3. The empirical validation of these metrics is shown in section 4. Finally, section 5 summarises the paper and draw our conclusions.

2 METRICS FOR REFERENTIAL INTEGRITY

Foreign key is one of the main concepts of the relational model. It can be defined as follows [5]: Let R2 be a base relation. Then a foreign key in R2 is a subset of the set of attributes of R2, say FK, such that:

- There exists a base relation R1 (R1 and R2 not neccesarily distincts) with a candidate key CK, and

[†] This research is part of the MANTICA project (a proposal to CICYT-FEDER) and is partially supported by the MANTEMA project carried out by ATOS ODS, S.A. (ATYCA Dirección General de Tecnología y Seguridad Industrial del Ministerio de Industria y Energía).

- For all time, each value of FK in the current value of R2 is identical to the value of CK in some tuple in the current value of R1.

Related with the foreign key concept, the relational model includes the referential integrity rule: the database must not contain any unmatched foreign key values [5].

We propose two different metrics based on the referential integrity for measuring relational database schemas length and complexity.

Length Metric

Relational database schema length can be measured using the TDRT (Tables Depth Referential Tree) metric. This metric is based on [4] DIT metric (Depth Inheritance Tree, the maximum depth in the inheritance tree). Based on the referential integrity relation, TDRT is defined as the maximum number of levels of referential integrity among tables.

Complexity Metric

Referentiability degree (RD) of a scheme is defined as the number of foreign keys (NFK) of all the tables in the schema.

$$RD = \sum_{i=1}^{NT} NFK$$

Being NT the number of tables in the schema

Figure 1 shows an example of database schema. In this case TDRT=4 and RD = 8.

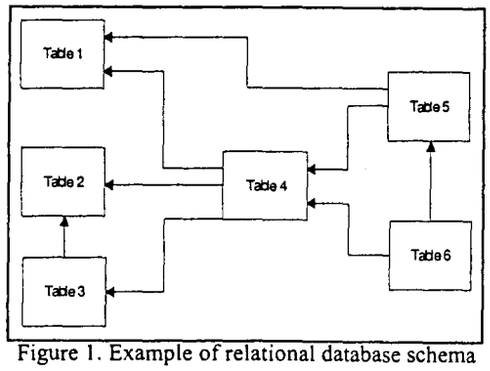

Figure 1. Example of relational database schema

3. METRICS FORMAL VERIFICATION

In this section we use the properties proposed by [3] in order to characterise the metrics defined in the previous section.

From a formal point of view, a relational scheme (S) is characterised by the set of its elements (E) which correspond to the set of the tables of a scheme with its referential integrity relations (R).

$$S=<E,R>$$

TDRT Metric

For length metrics, the properties given by [3] are:

1. Nonnegativity. The length of a system $S=<E,R>$ is nonnegative: $Length(S)\geq 0$
   TDRT fulfils this property because it never can be negative.

2. Null value. The size of a system S is null if E is empty. $(E=\varnothing) \Rightarrow Length(S)=0$
   If there are no tables $(E=\varnothing)$, TDRT=0.

3. Nonincreasing monotonicity for connected components. Let S be a schema and m a module of S such that m is represented by a connected component of the graph representing S (a set of tables related by referential integrity relations). Adding referential integrity relations between tables does not increase the length of S.

   $(S=<E,R>$ and $m=<E_m,R_m>$ and $m\subseteq S$ and m'' is a connected component of S'' and $S'=<E,R'>$ and $R'=R\cup \{<e_1,e_2>\}$ and $<e_1,e_2>\notin R$ and $e_1\in E_{m1}$ and $e_2\in E_{m1}) \Rightarrow Length(S)\geq Length(S')$

4. Nondecreasing monotonicity for nonconnected components. Let S be a system and m_1 and m_2 be two modules of S such that m_1 and m_2 are represented by two separate connected components of the graph representing S. Adding referential integrity

relations from tables of m_1 to tables of m_2 does not decrease the length of S.

($S=<E,R>$ and $m_1=<E_{m1},R_{m1}>$ and $m_2=<E_{m2},R_{m2}>$ and $m_1 \subseteq S$ and $m_2 \subseteq S''$ are separate connected components of S'' and $S'=<E,R'>$ and $R'=R \cup \{<e_1,e_2>\}$ and $<e_1,e_2> \notin R$ and $e_1 \in E_{m1}$ and $e_2 \in E_{m2}$) $\Rightarrow$ $Length(S') \geq Length(S)$

5. Disjoint modules. The length of a system $S=<E,R>$ made of two disjoint modules m_1, m_2 is equal to the maximum of the lengths of m_1 and m_2.

   ($S=m_1 \cup m_2$ and $m_1 \cap m_2=\emptyset$ and $E=E_{m1} \cup E_{m2}$) $\Rightarrow$ $Length(S) = \max\{Length(m_1), Length(m_2)\}$

   By definition, TDRT fulfils this property

RD Metric

To demonstrate that RD is a complexity metric, we prove that verifies the properties given by [3] for this kind of metrics:

1. Nonnegativity. The complexity of a system $S=<E,R>$ is nonnegative. $Complexity(S) \geq 0$

2. Null value. The complexity of a schema is null if it has no referential integrity relations. $(R=\emptyset) \Rightarrow (Complexity(S)=0)$

3. Symmetry. The complexity of a schema does not depend on the convention chosen to represent the referential integrity relations between its elements.

   ($S=<E,R>$ and $S^{-1}=<E,R^{-1}>$) $\Rightarrow$ $Complexity(S) = Complexity(S^{-1})$

   The definition of RD is the same disregarding the direction of the reference.

4. Module Monotonicity. The complexity of a schema $S=<E,R>$ is no less than the sum of the complexities of any two of its modules with no referential integrity relationships in common.
   ($S=<E,R>$ and $m_1=<E_{m1},R_{m1}>$ and $m_2=<E_{m2},R_{m2}>$ and $m_1 \cup m_2 \subseteq S$ and $R_{m1} \cap R_{m2}=\emptyset$) $\Rightarrow$ $Complexity(S) \geq Complexity(m_1)+Complexity(m_2)$

   If the modules are no disjoint, this means that between elements of both modules there is a relation of referential integrity, so RD never decrease.

5. Disjoint Module Additivity. The complexity of a schema composed of two disjoints modules is equal to the sum of the complexities of the two modules

   ($S=<E,R>$ and $S= m_1 \cup m_2$ and $m_1 \cap m_2=\emptyset$) $\Rightarrow$ $Complexity(S) \geq Complexity(m_1) + Complexity(m_2)$

   Every module will have a value for NFK. When modules are disjoint neither foreign key nor a table will be common to both modules, so the result of RD of the system will be the sum of the NFK of the two modules, and so RD will be the sum of the RD of the modules.

4 EMPIRICAL VALIDATION FOR THE METRICS

Besides the formal characterisation of the proposed metrics, an empirical validation has been carried out following the experimental method applied to software engineering ([11], [1]).

Our objective is demonstrate that the proposed metrics can be used **for measuring** the complexity of the relational database schema which influences in its understandability.

Hypotheses

The formal hypotheses are:

- Null hypothesis: Different values of metrics do not affect the comprehension of the database schema.

- Alternative hypothesis 1: The value of the TDRT metric affects the comprehension of the database schema.

- Alternative hypothesis 2: The value of the RD metric affects the comprehension of the database schema.

- Alternative hypothesis 3: The combination of the TDRT and RD metrics affects the comprehension of the database schema.

Subjects

The participants of the experiment are Computer Science students at the University of Castilla-La Mancha (Spain), who were enrolled in the databases course lasting two semesters.

Until the day of the experiment, the students did not known that they were to do it. The experiment was developed by 60 students but only 59 were finally accepted.

We have tried minimize variability among participants choosing the people of the same degree, the third (the last in Computer Science BSC). Effects of irrelevant variables were minimize making the same trials for all the subjects with the same duration (ten minutes per test).

Experimental materials

To test the hypotheses stated in the section 4.1. four separate software designs were required. In each one the values of the two metrics were different. There were two possible values for RD metric (eight or five) and for TDRT metric (two or five).

The documentation accompanying each design was approximately seven pages long and includes the schema database, the tables with their rows and the question/answer paper. For each design the database schema had six tables.

The subjects were asked to perform three tasks with the values of the database schema: insert, delete and update.

Figure 2 shows the question/answer paper associated to the database schema of the figure 1.

Before the experimental subjects make the test, the experiment was made on a small set of people in order to improve it.

1. What tables and how much rows in each table are affected if we delete in the Table 5 the row with cod1=210?

Table 1	Table 2	Table 3	Table 4	Table 5	Table 6

2. What tables and how much rows in each table are affected if we update the column X of the row with cod2=11 in the table 3?

Table 1	Table 2	Table 3	Table 4	Table 5	Table 6

3. What tables and how much rows and columns are necessary to add if we want add a new row in the table 4? (Suppose that all the necessary data are news in the database)

Table 1	Table 2	Table 3	Table 4	Table 5	Table 6

Experimental design

Each level of one factor appears with each level of the other one, so, we have selected the crossing design. This crossing relationship is denoted AxB. For us, A is the RD metric and B is the TDRT metric. See Table 1.

For increasing the power of the test, α has been set to 0.1 instead 0,05 level which is more common [2].

		Factor B (RD)	
		LOW	HIGH
Factor A (TDRT)	LOW	2,5	2,8
	HIGH	5,5	5,8

Table 1. Crossed Design for the experiment

Experimental results

There are three major items to consider when choosing the analysis techniques: the nature of the data collected, why is performed the experiment and the type of experimental design used [11].

Due to the type of the experiment used, F statistic is the most appropriate technique for obtain the results [14].

Table 2 shows the results for the F-statistic:

Source of Variation	Q_i	Degrees of Freedom	s_i^2	F-Ratio
TDRT	18.457	1	18.5	1.67
RD	531.000	1	531	48.1
Interaction	31.339	1	31.3	2.84
Error	2560.304	1	11.0	
Total	3141.102	232		

Table 2. Results of the F-statistic

Comparing these values with $F_{1,232}$=2.73, we can ensure that:

- Alternative Hypothesis 1: *"The value of the TDRT metric affect the comprehension of database schema"*

  Since 1.67<2.73, TDRT does not affect the results of the experiment

  So, alternative hypothesis 1 is invalid because the value of the TDRT metric do not affect the results obtained.

- Alternative Hypothesis 2: *"The value of the RD metric affect the comprehension of database schema"*

  Since 48.1>2.73, RD affects the results of the experiment

  So, alternative hypothesis 2 is valid because the value of the RD metric affects the results obtained.

- Alternative Hypothesis 3: *"The combination of the TDRT and RD metrics values affect the comprehension of database schema"*

  Since 2.84>2.73, the interaction of the metrics affects the results of the experiment

  So, alternative hypothesis 3 is valid because the combination of the values of the TDRT and the RD metrics affect the results obtained.

We can conclude that the number of foreign keys in a relational database schema is a more solid indicator of its understandability and that the **length of the** referential tree is no relevant by itself, but can modulates the effect of the number of foreign keys.

Folowing [15] RD can be classified as a dominant metric, and TDRT is not needed to classify quality (is a redundant metric).

CONCLUSIONS AND FUTURE WORK

There is a great need to measure the size, complexity and quality of databases[16].

We have proposed, verified and validated two metrics for measuring relational database complexity. These metrics are not sufficient to assess the quality of a relational database schema. Other metrics related to correctness, minimality, expressivity, cohesion etc must be elaborated.

However, the number of foreign keys of the schema (Referentiability Degree) has demonstrate to be a solid indicator of the complexity of the schema.

Now, we are elaborating other metrics for object-relational databases [12]. Some of them are metrics of length: *TDIT* that can be describe as the tables depth inheritance tree. Others are size metrics which measures the database scheme size.

REFERENCES

[1] Bourque, P., Côté, V.,"An Experiment in Software Sizing with Structured Analysis Metrics". *J. Systems Software.* 15:159-172. 1991.

[2] Briand,L., Bunse, C., Daly, J., Differding, C. "An experimental comparison of the Maintainability of Object-Oriented and Structured Design Documents", *Proc. Int. Conf. On Software Maintenance*, Harrold, M.J. Visaggid,G.(eds), Bari, 1-3 Oct, 130-138. 1997.

[3] Briand, L., Morasca, S. and Basili, V., "Property-Based Software Engineering Measurement". *IEEE Transactions on software Engineering,* Vol. 22, No 1, January. 1996

[4] Chidamber, S. and Kemerer, C., "A metrics suite for object-oriented design", *IEEE Trans. Software Eng.*, vol. 20,no. 6, June, pp. 476-493. 1994.

[5] Date, C., J., "An Introduction to Database Systems", 6th ed., Addison-Wesley. 1994.

[6] Frazer, A., "Reverse engineering-hype, hope or here?". In: P.A.V. Hall, Software Reuse and Reverse Engineering in Practice. Chapman & Hall. 1992.

[7] ISO. *Software Product Evaluation-Quality Characteristics and Guidelines for their Use.* ISO/IEC Standard 9126, Geneva. 1994.

[8] Li, H.F. and Cheng, W.K. "An empirical study of software metrics". IEEE Trans. on Software Engineering SE-13 (6), 679-708. 1984.

[9] McClure, C., "The Three R's of Software Automation: Re-engineering, Repository, Reusability". Englewood Cliffs: Prentice-Hall. 1992.

[10] Pearse, T., Oman, P., "Maintainability Measurements on Industrila Source Cade Maintenance Activities", Proceedings of the *International Conference on Software Maintenance,* Gianluigi Caldiera and Keith Bennett (eds), pp. 295-303. 1995.

[11] Pfleeger, S.L., "Experimental Design and Analysis in Software Engineering", Annals of Software Engineering1, pg. 219-253. J.C. Baltzer AG, Science Publishers. 1995.

[12] Piattini, M., Calero, C., Polo, M. and Ruiz, F. .Maintainability in object-relational databases. Proc. *of The European Software Measurement Conference FESMA'98*, Coombes, Hooft and Peeters (eds.), pp. 223-229. 1998.

[13] Pigoski, T.M., *Practical Software Maintenance. Best Practices for Managing Your Investment.* John Wiley & Sons. USA. 1996.

[14] Rohatgi, V.K., "An introduction to Probability Theory and Mathematical Statistics", *Wiley Series in Probability and Mathematical Statistics.* 1976.

[15] Schneidewind, N. F., "Software Metrics for Quality Control", Proceedings of the *Fourth International Software Metrics Symposium*, IEEE Computer Society Technical Council on Software Engineering, pp.127-136. 1997.

[16] Sneed,H. M., Foshag, O., "Measuring legacy database structures", proceedings of The European Software Measurement Conference, Technologisch Instituut, pp. 199-211. 1998.