JISBD 2004

Jornadas de Ingeniería del Software y Bases de Datos

Málaga, 10 · 12 Noviembre 2004

http://jisbd2004.lcc.uma.es

Juan Hernández

Ernesto Pimentel

Índice

Artículos cortos

Validating OCL Metrics through a Family of Experiments

Luis Reynoso[1], Marcela Genero[2], Mario Piattini[2]

[1] Department of Computer Science, University of Comahue,
Buenos Aires 1400, 8300, Neuquén, Argentina
lreynoso@uncoma.edu.ar
[2] Alarcos Research Group- Department of Computer Science,
University of Castilla-La Mancha.
Paseo de la Universidad, 4, 13071, Ciudad Real, Spain
{Marcela.Genero,Mario.Piattini}@uclm.es

Abstract. In recent years an important emphasis had been put on measuring structural properties of models specified by the Unified Modeling Language (UML) during the early stages of object-oriented (OO) development, but only focusing on the model elements expressed by diagrammatic notation. However diagram-based UML notation is limited in its expressiveness due to the fact that not all model constraints can be captured, producing thus a model that would be severely underspecified. This is why the Object Constraint Language (OCL) has emerged, allowing us to specify additional model constraints in a more precise and concise way than it is possible to do with diagrams only. The lack of metrics which capture quality aspects of OCL expressions motivated us to define a set of metrics for measuring their structural properties. The main goal of this paper is to carefully describe a family of experiments, which we have undertaken to ascertain if any relationship exists between two structural properties, length (measured by the Depth of Navigations metric, DN) and coupling (measured by the Number of Navigated Classes metric, NNC), and two maintainability sub-characteristics: understandability and modifiability of OCL expressions. Even though the results obtained show empirical evidence that such a relationship exists, they must be considered as preliminaries. Further validation is needed, including experimentation with practitioners and application of the metrics to "real projects".

1 Introduction

Most key development decisions are usually made at the early system specification stage of a software project [21]. As early decisions have a high impact on the quality of the OO software systems which are finally delivered, special attention has been paid to the quality of "early" artifacts. This fact is corroborated by the huge amount of metrics that have been proposed in the literature for measuring internal quality attributes of UML [16]. But those proposals of metrics only focus on the UML model elements expressed by diagrammatic notation [1], [5], [6], [10].

Even though UML has become the standard language of OO modeling [22], many design decisions, constraints and essential aspects of software systems cannot be expressed using only UML diagrammatic notation [22]. In order to alleviate this problem OCL [15] has emerged as a constraint language. With the introduction of OCL by OMG, the quality of early models can be improved specifying it in a combination of the UML and OCL languages, i.e., through UML/OCL combined models. Without OCL expressions the model would be severely underspecified [22]. This led us to think about the necessity of having metrics to measure structural properties for OCL expressions within UML/OCL models, with the idea that these metrics could be early indicators of OCL expressions understandability and modifiability.

When defining the OCL metrics we considered in turn the OCL concepts involved with "chunking" and "tracing" cognitive techniques [17], [18], trying to cover most of the concepts of the OCL metamodel [15]. These techniques are concurrently and synergistically applied in problem solving and were defined by Cant et al. [8], [9] in the Cognitive Complexity Model (CCM model). "Chunking" involves the recognition of a set of declarations and extracting information from them, which is remembered as a chunk (a single mental unit), whereas 'tracing' involves scanning, either forward or backwards, in order to identify relevant 'chunks'. The last mentioned technique has been observed as an essential activity in software comprehension [4].

As the metric definition and the theoretical validation were previously done in [17], [18] this paper focuses only on the empirical validation. For that reason, the main goal of this paper is to carefully describe a family of experiments we have undertaken to ascertain if any relationship exists between two metrics: a length metric, Depth of Navigation (DN), and coupling metric, Number of Navigated Classes (NNC), and two maintainability sub-characteristics [12]: understandability and modifiability of OCL expressions. The definition of both metrics is shown in Table 1.

Table1. Metric definition

Number of Navigated Classes (NNC)
Definition: This metric counts the total number of classifiers (classes, association classes, interfaces, etc.) to which an expression navigates to (these classes are frequently used in an expression through relationships role names). If a class contains a reflexive relation and an expression navigates it, the class will be considered only once in the metric. **Goal:** Warmer and Kleppe [22] argue that "any navigation that traverses the whole class model creates a coupling between the objects involved". A higher number of navigated classes will increase the coupling between the objects.
Depth of Navigations (DN)
Definition: Given that in an OCL expression there can be much navigation regarding its definition, we build a tree of navigation using the class names used in navigations. We will only consider navigations starting from the contextual instance (from self). The root of the tree is the class name which 'self' represents. Then we build a branch for each combined navigation, where each class we navigate to is a node in the branch. DN is defined as the maximum depth of the tree. **Goal:** A higher depth of navigations may involve a complicated navigation. Warmer et al. [22] suggest avoiding complex navigation expressions, and they also argue that: "using long navigation makes details of distant objects known to the object

where we started the navigation". It is based on the idea that a higher value of this metrics will be an indicator of how distant the objects known by the Classifier where the expression is defined are.

DN is related to the "tracing" technique due to its use of an important OCL concept: the navigation, whereas NNC is related to the "chunking" technique because it involves the comprehension of a class as a chunk. The fact that DN and NNC are related to length and coupling respectively can be found in their theoretical validation [17], [18], which was obtained following a property-based framework proposed by Briand et al. [6], [7]. Although the DN metric was validated as a length metric its meaning is closely related to coupling concepts as its goal describes.

Coupling is a more complex software attribute in OO systems, and there are many different mechanisms that can constitute coupling [5]. A high quality software design, among many other principles, should obey the principle of low coupling [5]. The metric DN reveals how dependent the contextual instance of distant objects in a class diagrams is, while NNC reveals how many different objects are coupled to the contextual instance.

In relation to our main goal the following section describes a family of experiments we carried out to empirically validate the aforementioned metrics. Finally, in section 3 some conclusions are drawn and future work is described.

2 Family of Experiments

As Miller [14] and Basili et al. [3], among others, suggested simple studies rarely provide definite answers. Following these suggestions, we have run an experiment an two replicas. Both replicas were based on the same hypotheses of the original experiment. As most of the details of the replicas were very similar to the experiment, we will first describe the common aspects of the family of experiments, and later on we will only focus on those characteristics that are different for each experiment.

2.1 Common aspects of the family of experiments

In order to conduct the experimental process we have followed (with minor changes) the format proposed by Wohlin et al.[23]. In this section we will summarize its main steps. Further information of the experiment, mainly its threats and laboratory package is available at http://alarcos.inf-cr.uclm.es.

Experiment goal[1]: Analyze <<Two factor length and coupling, measured by DN and NNC respectively>>; for the purpose of <<Evaluating>>; with respect to <<The capability to be used as understandability and modifiability indicators of OCL expressions>>; from the point of view of <<researchers>>: in the context of <<Undergraduate Computer Science students>>.

Variable selection. The independent variables were two structural properties: length and coupling. The dependent variables were two maintainability sub-characteristics:

[1] The goal is defined using the GQM [2] template for goal definition.

understandability (UND) and modifiability (MOD) of OCL expressions.

Instrumentation. The objects were four UML/OCL combined models according to the experimental design we had chosen, each of them having one OCL expression. These models were related to different universes of discourse that were easy enough to be understood by each of the subjects. Each model had a test enclosed that included two types of tasks:

- UND tasks: These consist of four questions about the meaning of the OCL expression within the UML/OCL model. These questions reflected whether or not the experimental subject had understood each diagram. The subject also had to note how long it took to answer the questions.
- MOD tasks: Each subject had to modify the OCL expression according to three new requirements. The modifications to each test were similar, including defining new navigations, attributes referred through navigations, etc.

The independent variable was measured through the DN and the NNC metrics. The dependent variables were measured according to the number of seconds the subject spent performing the UND tasks and the number of seconds the subject spent performing the MOD tasks.

We also design a debriefing questionnaire including personal details and experience of the subjects which help us to identify their profile.

Hypotheses formulation. We wish to test the following hypotheses:

- Null hypothesis, $H_{0,1}$: There is no effect of length (measured by DN) on the UND time of OCL expressions. // $H_{3,1}$: $\neg H_{0,1}$
- Null hypothesis, $H_{1,1}$: There is no effect of coupling (measured by NNC) on UND time of OCL expressions. // $H_{4,1}$: $\neg H_{1,1}$
- Null hypothesis $H_{2,1}$: There is no interaction effect between length (measured by DN) and coupling (measured by NNC) on UND time of OCL expressions. // $H_{5,1}$: $\neg H_{2,1}$

Analogously hypotheses for the MOD time are defined, named as $H_{i,2}$, i=0..5.

Experiment design. Taking the hypotheses into account, we considered two factors: NNC and DN metrics, having each two values (low, high), 1 and 3 for DN, and 2 and 4 for NNC. The 2x2 crossed factorial design is shown in Table 2. We selected a within-subject design experiment, i.e., all the tests (experimental tasks) had to be solved by each of the subjects. The tests were put in a different order for each subject for alleviating learning effects.

Table 2. A 2X2 crossed factorial design

		DN	
		Low	High
NNC	Low	2,1 (G1)	2,3 (G3)
	High	4,1 (G2)	4,3 (G4)

Data validation. In order to check if the collected data is correct the UND correctness (Number of correct answers /Number of questions answered) and MOD correctness (Number of correct modifications/Number of modifications applied) indicators were used to validate the data.

We used box-plot in order to identify any possible outliers, representing unusual data points in the collected data due to an excessive time of subjects performing the UND

or MOD tasks (usually they correspond to few subjects with learning or fatigue problems).
Analysis and Interpretation[2]. Before testing the formulated hypothesis we evaluated if the data follow a normal distribution or not using the Shapiro–Wilk test. We decided to select α =0.05 which means a 95% level of confidence. Due the data for the experiment and its two replicas was normal, we decided to carry out an ANOVA with repeated measures [20], because this type of analysis allows us to analyze interaction between the independent variables under study and the measurement of the dependent variable is repeated.

2.2 Description of the original experiment

The main characteristics of the original experiment, which was carried out in Neuquén (Argentina), are:
Subjects. In order to select the subjects we motivated a group of students who had taken a semester on System Analysis (at the National University of Comahue, UNC, Argentina) to take an additional and intensive course about OCL language. In the last lecture of the course where the experiment was run, fifteen students participated. The profile of the subject was the following: their average age is 24 years old, they have an average of 4 years of experience in programming, and one year in modeling with UML.
Data validation. After we collected the data we found all data was correct. However, in the understandability part we decided to discard 20 tests (of 5 subjects) whose UND correctness was lower than 0.75 (only 2 out of 4 questions were correctly answered in all the discarded tests). We think it was reasonable to discard these tests because the level of correctness is important in order to obtain reliable and relevant results about understandability aspects. Analyzing the data obtained, one outlier was identified. So, 36 tests of the UND tasks were analyzed. Regarding the MOD tasks, the tests belonging to one subject was discarded due their correctness was lower than 0.75. In this case, three outliers were identified. So, 44 tests were analyzed.
Analysis and Interpretation. Through the ANOVA we obtained that the DN metric has a great influence on the UND time and MOD time of OCL expressions. These findings revel, to some extent, that OCL expressions that have in their definition "longer" navigations take more time to understand and more time to modify than others having "shorter" ones.

2.3 First Replica

The main characteristics of the first replica carried out in Ciudad Real (Spain) are:

[2] All the data analysis was carried out by means of SPSS [20].

Subjects. The subjects were twelve undergraduate students enrolled in the fifth-year of Computer Science in the Department of Computer Science at the University of Castilla-La Mancha (UCLM) in Spain. The experiment was run in a practical session of the "Software Engineering II" course. The profile of the subjects was the following: their average age was 24 years old, they had an average of 4 and a half years of experience in programming, and two years in modeling with UML.
Data Validation. After collecting the data we checked if the tests were complete and if the modifications were well done. Regarding the UND tasks, we discarded the tests of five subjects: four subjects had at least one test with UND correctness less than 0.75, and one subject was found as an outlier. In the case of the MOD tasks, we discarded five subjects because three of them had at least one test with low correctness and two subjects were found as outliers. Therefore, we took into account the responses of 7 subjects for each part.
Analysis and Interpretation. By carrying out an ANOVA with repeated measures we found the value of the DN metric affects the UND time of OCL expressions within UML/OCL models, and the MOD time was affected by the NNC metric. ic.

2.4 Second Replica

The main characteristics of the second replica carried out in Alicante (Spain) are:
Subjects. The subjects were twenty nine students enrolled in the third-year of Computer Science in the Department of Computer Science at the University of Alicante (UA), Spain. They were students in the first Software Engineering course. We invited the students to do a short seminar about OCL and to do a test as part of the seminar. The subjects were motivated to participate. The profile of the subjects is the following: their average age is 22 years old, they have an average of 2 years experience in programming, and less than one year in modeling UML class diagrams.
Data Validation. In the understandability part, the tests of thirteen subjects, with a correctness less than 0.75 were discarded. A similar number of subjects, fourteen, were discarded in the MOD tasks. Two outliers were identified in the MOD tasks.
Analysis and Interpretation. The results obtained through the ANOVA with repeated measures show the interaction of DN and NNC metrics affects the UND time of OCL expression and the DN metric influence on the MOD time.

3 Conclusions and Future Work

Even though applying OCL to software specification has great potential for improving software quality and software correctness [11], there are no metrics that capture structural properties of OCL expressions. We are aware that it is necessary to define metrics for OCL expressions since OCL is a textual language that allows us to specify additional constraints on models in a more precise and concise way than it is possible to do with diagrams only. For that reason we have previously defined a theoretically validated a set of metrics for the structural properties of OCL expressions [17], [18].

As many authors mentioned [2], [10], [13], [19] empirical validation of metrics, through experiments or case studies is fundamental to assure that the metrics are really significant and useful in practice. For that reason in this paper, we have presented a controlled experiment and two replicas for assessing if two of the metrics we proposed are related to OCL expression's understandability and modifiability. These metrics are Depth of Navigation (DN) and the Number of Navigated Classes (NNC).

Even tough the results obtained through the family of experiments are not identical (see summary results in Table 3) we can conclude that the values the Depth of Navigation have a great influence in the time the subjects spent on understanding and modifying OCL expressions. So, DN emerges as the most important indicator for both dependent variables. These reveals, that OCL expression understandability and modifiability are more dependent on how far distant objects are coupled to the contextual instance (DN) rather than how many different objects are coupled to it (NNC).

Nevertheless, these findings must be considered as preliminaries. As a future work, we will perform a replication of this experiment with practitioners in order to increase the external validity of the results, and we will also use the metrics in "real projects". We also plan to use in future experiments, UML/OCL models of higher size. The empirical validation of the rest of metrics is also pending.

Table 3. Summary of the experiment and their replicas results

	Experiment in Neuquén	**Experiment in Ciudad Real**	**Experiment in Alicante**
Understandability Time	Affected by DN	Affected by DN	Affected by the interaction of DN and NNC
Modifiability Time	Affected by DN	Affected by NNC	Affected by DN

Acknowledgements

This research is part of the MESSENGER project (PCC-03-003-1) financed by "Consejería de Ciencia y Tecnología de la Junta de Comunidades de Castilla-La Mancha (Spain)", the CALIPO project supported by "Dirección General de Investigación del Ministerio de Ciencia y Tecnologia (Spain)" (TIC2003-07804-C05-03), the network VII-J-RITOS2 financed by CYTED, and the UNComa 04/E048 project financed by "Universidad Nacional del Comahue". Luis Reynoso enjoys a postgraduate grant from the agreement between the Government of Neuquen Province (Argentina) and YPF-Repsol. We will also thank to Cristina Cachero at the Alicante's University and Macario Polo at the Castilla La Mancha's University for their cooperation for running this experiment.

References

1. Bansiya, J., Davis, C.G.: A Hierarchical Model for Object-Oriented Design Quality Assessment, IEEE Transactions on Software Engineering, Vol. 28 Nº 1, (2002) 4-17

2. Basili, V. R., Rombach, H.D.. The TAME project: towards improvement-oriented software environments. IEEE Transactions on Software Engineering. Vol. 14 N° 6 (1998) 758-773
3. Basili, V.R., Shull, F., Lanubile, F.: Building knowledge through families of experiments. IEEE Transactions on Software Engineering, Vol. 25 N° 4 (1999) 456-473
4. Boehm-Davis, D.A., Fox, J.E., Philips, B.: Techniques for Exploring Program Comprehension. Empirical Studies of Programmers. Sixth Workshop. Eds. W. Gray and D. Boehm-
5. Briand L.C., Bunse, C., Daly, J.W.: A Controlled Experiment for evaluating Quality Guidelines on the Maintainability of Object-Oriented Designs. IEEE Transactions on Software Engineering, Vol. 27 N° 6 (2001) 513-530
6. Briand, L.C., Morasca, S., Basili, V.R.: Defining and Validating Measures for Object-based High Level Design. IEEE Transactions on Software Engineering. Vol. 25 N° 5, (1999) 722-743
7. Briand, L.C., Morasca, S., Basili, V.R.: Property-based Software Engineering Measurement. IEEE Transactions on Software Engineering, Vol. 22 N° 1 (1996) 68-86
8. Cant, S.N., Henderson-Sellers, B., Jeffery, D.R.: Application of Cognitive Complexity Metrics to Object-Oriented Programs. Journal of Object-Oriented Programming, Vol. 7 N° 4 (1994) 52-63
9. Cant, S.N., Jeffery, D.R., Henderson-Seller, B.: A Conceptual Model of Cognitive Complexity of Elements of the Programming Process. Information and Software Technology, Vol. 37 N° 7 (1995) 351-362
10. Fenton, N.E., Pfleeger, S.L.: Software Metrics: A Rigorous and Practical Approach. Chapman & Hall, London, 2nd Edition. International Thomson Publishing Inc. (1997)
11. Hennicker, R., Hussmann, H., Bidoit, M.: On the Precise Meaning of OCL Constraints. Tony Clark and Jos Warmer, editors, Advances in Object Modelling with the OCL, Springer, Berlin, LNCS 2263 (2001) 69-84
12. ISO/IEC 9126. Software Product Evaluation-Quality Characteristics and Guidelines for their Use. Geneva.
13. Kitchenham, B., Pflegger, S., Fenton, N.: Towards a Framework for Software Measurement Validation. IEEE Transactions of Software Engineering, Vol. 21 N° 12 (1995) 929-944
14. Miller, J.: Applying Meta-Analytical Procedures to Software Engineering Experiments, Journal of Systems and Software, 54 (2000) 29-39
15. Object Management Group. UML 2.0 OCL 2nd revised submission. OMG Document ptc/03-10-14. [On-line] Available: http://www.omg.org/cgi-bin/doc?ptc/2003-10-14
16. Object Management Group. UML Specification Version 1.5, OMG Document formal/03-03-01. [On-line] Available: http://www.omg.org/cgi-bin/doc?formal/03-03-01
17. Reynoso, L., Genero, M., Piattini, M.: Measuring OCL expressions: a "tracing"-based approach. Workshop on Quantitative Approaches in Object-Oriented Software Engineering. QAOOSE (2003) Germany
18. Reynoso, L., Genero, M., Piattini, M.: Measuring OCL Expressions: An approach based on Cognitive Techniques. Piattini M., Genero M. and Calero C. Editors, Imperial College Press, UK (to appear), 2004
19. Schneidewind, N.F.: Methodology For Validating Software Metrics. IEEE Transactions of Software Engineering, Vol. 18 N° 5 (1992) 410-422
20. SPSS, 2002 SPSS 11.5. "Syntax Reference Guide". Chicago. SPSS Inc. (2002)
21. Vinter, R., Loomes, M., Kornbrot, D.: Applying Software Metrics to Formal Specifications: A Cognitive Approach. 5th. Int.Symposium on Software Metrics (1998) 216-223
22. Warmer, J., Kleppe, A.: The Object Constraint Language. 2nd edn. Getting Your Models Ready for MDA. Object Technology Series. Addison-Wesley, Massachusetts (2003)
23. Wohlin, C., Runeson, P., Höst, M., Ohlson, M., Regnell, B., Wesslén, A.: Experimentation in Software Engineering: An Introduction, Kluwer Academic Publishers (2000)