2003

Evaluation of the Impacts of Data Model and Query Language on Query Performance

Hock Chuan Chan
National University of Singapore, chanhc@comp.nus.edu.sg

Lian Xiang
National University of Singapore, g0201698@nus.edu.sg

Evaluation of the Impacts of Data Model and Query Language on Query Performance

Hock Chuan Chan
National University of Singapore
chanhc@comp.nus.edu.sg

Lian Xiang
National University of Singapore
g0201698@nus.edu.sg

ABSTRACT

It is important to understand how users can utilize database systems more effectively to enhance performance. A major research interest is to evaluate and compare user performance across different data models and query languages. So far, experiments have tested combinations of model plus language. An interesting theoretical and practical question is: how much of the performance difference is caused by the data model itself, and how much by the additional query language syntax? A cognitive model of query processing suggests measurement at two stages. The data model has impact at the first stage, and the model with the query language syntax together has the impact at the second stage. An experiment that compares the objected-oriented and relational models and query languages at the two stages provides fresh results.

Keywords

Data model, query language, user performance, empirical study, query stage

INTRODUCTION

Databases form an integral part of organizational systems. The evaluation and explanation of how users can make effective use of databases is an important area of information systems research, which has seen a steady stream of empirical studies (Aversano et al., 2002; Bowen and Rohde, 2002; Borthick et al., 2001; Chan et al., 1999; Siau et al., 1997; Owei and Navathe, 2001).

Many studies have been done on relative comparison of data models and query languages. For experiment studies on modeling performance, there is one main database variable: the data model. Differences in modeling performance can be readily attributed to the model. For studies on query performance, the main database variable is a combination of a data model and a query language. Studies have typically required subjects to write queries. The process involves a combination of data model and query language knowledge. So far, differences in user query performance have been attributed to the combination of data model and query language. Findings in the literature do not show whether the data model or the query language has more impact on query performance, leaving a lingering doubt on the interpretation and even validity of the findings.

This study addresses this issue in a comparison of the objected-oriented and relational models. It compares the user performance differences because of the impact of data model itself, and also compares the differences because of the additional impact of a query language within a model. Section 2 presents a cognitive model of the query process, which allows us to measure the effect of the model alone and the effect of the model plus query syntax. Section 3 presents the research methodology, followed by the results of the experiment. Lastly the conclusion is given.

A COGNITIVE MODEL OF DATABASE QUERY

This section provides a cognitive perspective on how data model and query language influence query performance. Ogden (1985) proposes a three-stage cognitive model of database query: query formulation stage (stage 0), query translation stage (stage 1), query writing stage (stage 2). The model is illustrated in Figure 1. The query formulation stage is concerned about real world data. An example is "Who are the faculty members in the business school?"

Based on the question from stage 0, users decide what elements of the data model are relevant, and the necessary operations. This is the query translation stage. For example, the output of this stage is "The faculty relation is needed, the column name is to be selected, and there is a condition for school name to be 'business'". This output need not be written down. In the query writing stage, users have to specify the output from stage 1 into the formal syntax of a query language. A simple example in SQL is: "select name from faculty".

There are many other models that involve similar steps in the query process. For example, the model by Mannino (2001) has two steps: from problem statement to database representation, and from the database representation into a database query language statement. Reisner (1977) proposes a process where a user will generate a set of lexical items and also generate a query template, followed by the merging of the lexical items with the template to generate the final query. The correspondence to the query translation and query writing stages are clear. This model is also related to the idea of semantic and articulatory distances as used in Liao and Palvia (2000). The articulatory distance is about stage 2, where users need to articulate the answers in a formal syntax. The semantic distance is about stage 1.

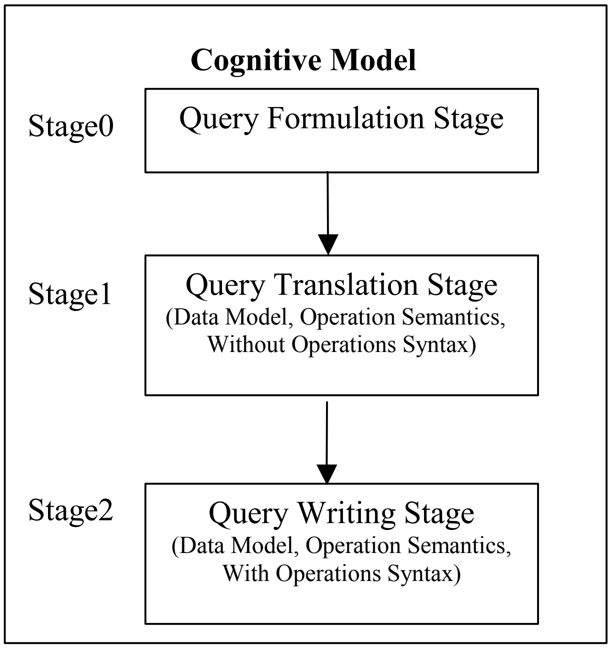
Cognitive Model
Stage0
Query Formulation Stage
Stage1
Query Translation Stage
(Data Model, Operation Semantics, Without Operations Syntax)
Stage2
Query Writing Stage
(Data Model, Operation Semantics, With Operations Syntax)

Figure 1. Query Model

Prior experiments on query performance have measured user performance after stage 2. If we can measure user performance after stage 1, and after stage 2, it will be possible to have a better understanding of the relative impact of model alone and the model with the additional language syntax.

RESEARCH MODEL AND METHODOLOGY

Research Model and Variables

Performance is influenced by four major factors: data-model/query language, task, user and system characteristics (Reisner, 1981; Chan et al., 1993). The independent variable is the abstraction level of the data model, set at two levels: the conceptual level where subjects used a version of OO model (O_2) with OQL, and the logical level where subjects used the relational model with SQL. The research model highlighting the comparison within stages (across models with / without query language) and comparison across stages (within the object-oriented model or within the relational model) is shown in Figure 2.

There have already been many empirical studies on the effects of data models and query languages (Liao and Palvia, 2000; Chan et al., 1993; Wu et al., 1994), which suggest that the conceptual level models (OO and ER) will lead to better user performance than the logical level model (relational), at least for the query writing stage. The different performance between the abstraction levels has been attributed to the type and amount of knowledge. For example, at the conceptual level, the objects such as entities and relationships are closer to the real world semantics which users are familiar with. On the other hand, at the logical level, the constructs are relations and primary keys / foreign keys which users are not familiar with. With ideal implementations, a relationship at the conceptual level can be specified quite easily (e.g. employee.department in a typical object-oriented query), compared to the unfamiliar specification of joins at the logical level (e.g. employee.eno. = department.empno in an SQL query). A more detailed description of the abstraction levels can be found in Chan et al. (1993).

So far, there are no studies that measure user query performance at the two different stages. We make the following hypothesis:

H1: Subjects using O_2 / OQL will perform better (in terms of accuracy, time and confidence) than subjects using relational / SQL for the query writing stage.

H2: Subjects using O_2 / OQL will perform better than subjects using relational / SQL for the query translation stage.

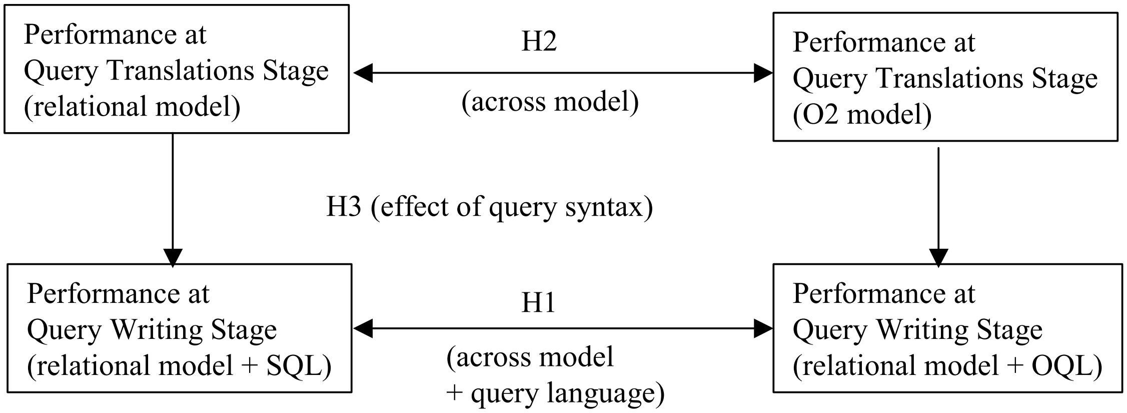
Performance at Query Translations Stage (relational model)
H2
(across model)
Performance at Query Translations Stage (O2 model)
H3 (effect of query syntax)
Performance at Query Writing Stage (relational model + SQL)
H1
(across model + query language)
Performance at Query Writing Stage (relational model + OQL)

Figure 2. The Research Model (Performance is measured by accuracy, time taken and confidence)

H3: Performance will be better at the query translation stage than at the query writing stage, for both groups.

The third hypothesis compares two different tasks. The difference can be attributed to the effect of the additional query language syntax.

Research Method and Process

A laboratory experiment was conducted to test the hypotheses. Each subject performed eight questions for both stages. These queries covered the basic queries that are commonly made on the relational model. Extraneous factors are controlled through randomization (for individual characteristics) or through standardization across groups (for interface characteristics).

Subjects were trained before they took the query test. The program displayed the questions one by one. They first finished query translation and then the query writing for each query. Each subject was given a relational schema or a diagram of an OO model, on paper. The test materials are in the Appendix. The query answers, the time taken in seconds, and the confidence level for each query were recorded by the computer.

EXPERIMENT RESULT AND DISCUSSION

The mean and standard deviation (in parenthesis) for accuracy, time, and confidence are shown in Table 1. Since the stage 1 data do not follow a normal distribution, non-parametric tests, using SPSS, were used. The Mann-Whitney independent sample test is used to compare between groups. The results show that the OO group is significantly more accurate than the relational group for stage 1 (z=-4.09, p=0.001) and stage 2 (z=-4.66, p=0.001). Time and confidence do not show significant differences. Thus, hypotheses 1 and 2 are both supported for accuracy measure, and not for time and confidence measures. This result corroborates previous studies for stage 2 (Chan et al., 1993; Wu et al., 1994), and provides new evidence for stage 1 differences.

	Relational Model		OO Model	
	Query Translation	Query Writing	Query Translation	Query Writing
Accuracy	4.58 (.43)	3.31 (.53)	4.85 (.29)	4.38 (.56)
Time	50.5 (21.5)	169.5 (47.3)	50.1 (16.4)	146.7 (45.4)
Confidence	4.80 (.53)	4.17 (.72)	4.81 (.36)	4.25 (.77)

Table 1. Mean (Standard Deviation) of Measures

Table 2 shows the results across query stages, using non-parametric Wilcoxon signed ranks test. Both groups show better performance at the query translation stage than at the query writing stage, for all measures of accuracy, time and confidence. Hypothesis 3 is fully supported. This shows that the query language syntax imposes significant additional difficulty to the query process. Furthermore, we find that many subjects with fully correct answers in stage 1 made serious mistakes in stage 2. Thus, even when subjects fully know what they want (the data structures and operations in query translation stage), they have difficulties putting that in the formal syntax required by a query language. These findings apply to both the relational and OO groups.

Data Model	Accuracy	Time	Confidence

Relational Model	z=-3.921[a]
	p=0.000**
		z=-3.920[b]
		p=0.000**
			z=-3.935[a]
			p=0.000**

OO Model	z=-2.952[a]
	p=0.003*
		z=-3.920[b]
		p=0.000**
			z=-3.525[a]
			p=0.000**

a Based on positive ranks. b Based on negative ranks. * Significant at p<0.05 ** Significant at p<0.01			

Table2. Non-parametric Test across Stages

Figure 3 illustrates accuracy performance at different stages of the cognitive model. At stage 0, we assume that the subjects can understand the meaning of query questions (and so a value of 5 is given). At this point, we are able to return to the questions posed earlier.

1. How much of the overall drop in performance (from the ideal top score) can we attribute to the data model alone, and how much to the particular query language within a model? At stage 1, performance shows a slight drop from stage 0 (9% for relational model, and 3% for OO model). At stage 2, performance drops by a very large amount (28% for SQL, and 10% for OQL) compared to stage 1. These numbers indicate the relative difficulties imposed on the users by the data model, and by the query language (additional to the model). The syntactical requirements of SQL with relational model and OQL with OO model cause about 3 times the difficulties caused by the data model alone. What we see here is that users basically do know what they want (and they can even perform the operations mentally to identify the right data values, on a small data set), but they have difficulties expressing them in a formal query language.
2. The OO model leads to better query results than the relational model, supporting findings in the literature. How much of this difference can be attributed to the models, and how much to the languages? This study shows that models alone cause a small 0.27 (out of 5) difference in accuracy. But at stage 2, when the query languages have been added to the data model, the difference is much bigger: 1.07. Thus, only about one third of the overall difference across models/query languages can be attributed to the models, and the other two thirds to the languages. This leads to the third question.

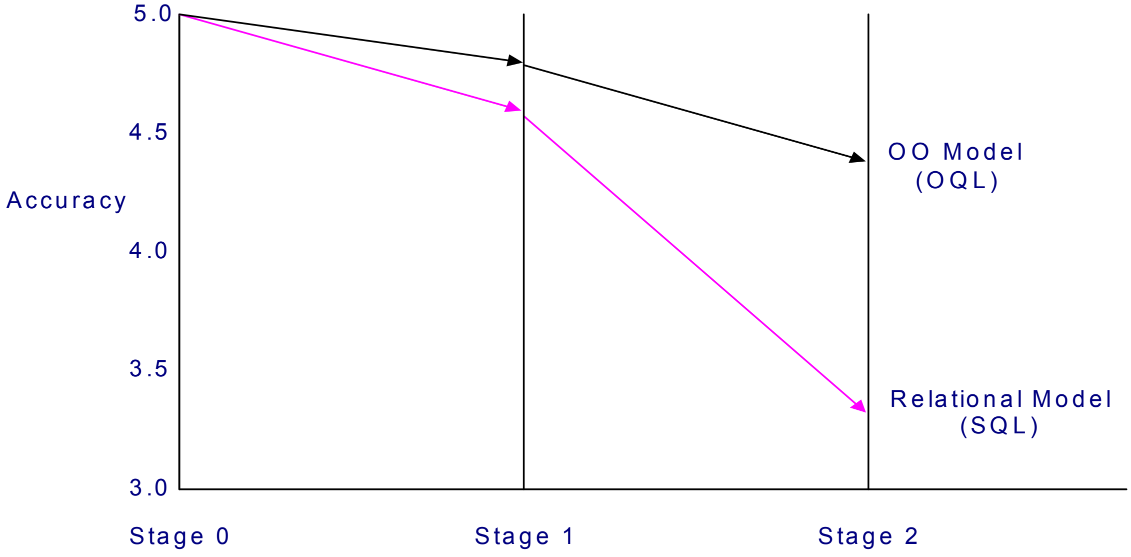

Figure 3. Query Performance at Each Query Stage

3. This study and others in the literature show a consistent finding that the OO/ER models are better than the relational model for query performance. But one doubt that is raised from figure 3 is: instead of using SQL, can we get a better relational language such that the overall query performance across models + languages will show no difference? We note that the relational model's performance at stage 1 is higher than the OO model's performance at stage 2. If a good language with little syntax difficulty can be found for the relational model, it could be possible that the overall query writing performance will show no difference. This is a challenge for researchers to develop a more user friendly relational textual query language.

CONCLUSION AND FUTURE WORK

Our experiment illustrates a finer approach to measure user query performance, based on a 3-stage cognitive model of query processing. By measuring query performance at different stages of the query process, we demonstrated the impacts of data models alone, for the object-oriented and the relational models, and the additional impacts of the query languages. The study shows that a higher abstraction level model leads to higher user performance for both query stages. The study also shows that the data model itself has a relatively small impact (about a third), and the query language has the remaining two thirds. It shows that generally users do know what they want, but have difficulty expressing it in a formal query language.

ACKNOWLEDGEMENT

We thank all the reviewers for their valuable comments.

REFERENCES

1. Aversano, L., G. Canfora, A. De Lucia, S. Stefanucci (2002). "Understanding SQL through iconic interfaces," *Proceedings 26th Annual International Computer Software and Applications Conference*, 703-708.
2. Borthick, A.F., P. L. Bowen, D. R. Jones and M. H. K. Tse (2001). "The effects of information request ambiguity and construct incongruence on query development," *Decision Support Systems*, 32(1) 3-25.
3. Bowen, P. L., F. H. Rohde (2002). "Further evidence of the effects of normalization on end-user query errors: an experimental evaluation," *International Journal of Accounting Information Systems*, 3(4) 255-290.
4. Chan, H.C., Wei, K.K. and Siau, K.L. (1993). "User-Database Interface: The Effect of Abstraction Levels on Query Performance," *MIS Quarterly*, 17(4) 441-464.
5. Chan, H. C., Tan, B. C. Y. and K. K. Wei (1999). "Three Important Determinants of User Performance for Database Retrieval," *International Journal of Human-Computer Studies*, 51(5), 895-918.
6. Liao, C. and P. C. Palvia, (2000), "The Impact of Data Models and Task Complexity on End-User Performance: an Experimental Investigation", *International Journal of Human-Computer Studies,* 52(5) 831-845.
7. Mannino, M.V. (2001). *Database Application*

Development and Design, McGraw-Hill Company, Inc.

8. Ogden, W.C. (1985). “Implications of a Cognitive Model of Database Query: Comparison of a Natural Language, a Formal Language, and Direct Manipulation Interface”, *ACM SIGCHI Bulletin,* 18(2) 51-54.
9. Owei, V., S. B. Navathe (2001) “Enriching the conceptual basis for query formulation through relationship semantics in databases,” *Information Systems*, 26(6) 445-475.
10. Reisner, P. (1977). “Use of psychological experimentation as an aid to development of a query language,” *IEEE Transactions on Software Engineering*, SE-3(3), 218-229.
11. Reisner, P. (1981). “Human Factors Studies of Database Query Languages: A Survey and Assessment.” *ACM Computing Surveys.* 13(1) 13-31.
12. Siau, K., Y. Wand, I. Benbasat, (1997). “The relative importance of structural constraints and surface semantics in information modeling,” *Information Systems*, 22(2/3) 155-170.
13. Wu, C.Z., Chan, H.C., Teo, H.H. and Wei, K.K. (1994). “An Experimental Study of Object-Oriented Query Language and Relational Query Language for Novice Users,” *Journal of Database Management,* 5(4) 16-27.

APPENDIX: DATABASE AND QUERIES FOR THE EXPERIMENT

This appendix contains the relational schema and the OO model, and the set of questions used in the experiments.

Query Questions:

1. Show the department name and city.
2. Show the engineers’ name and professions.
3. Show the names of employees who head any project.
4. Show the names of employees who work in the sales department.
5. Show the names of employees who work in the same department as Jack.
6. Show the names of employees with higher salaries than Jack’s.
7. List the names of managers who manage more than one department.
8. List the names of engineers who do not head any project.

Data Models:

Employee (number, name, salary)

Engineer (number, profession)

Manager (number, rank)

Department (number, name, city)

Project (number, name)

Work (employee number, department number, date)

Figure A1. The Relational Schema

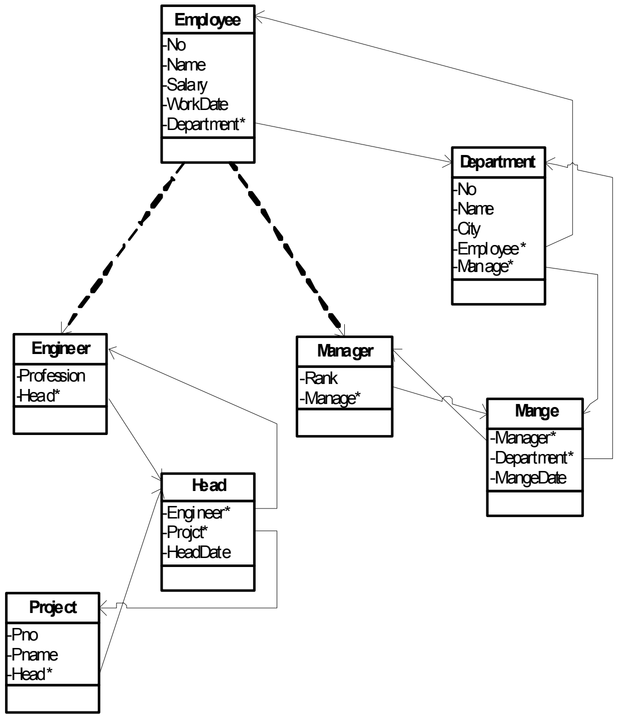

Figure A2. The Object-Oriented Data Model