

**A SYSTEMS AND PROCESS MODEL
FOR DATA EXPLORATION**

BY

JOHN PETER LEE

ABSTRACT OF A DISSERTATION SUBMITTED TO THE FACULTY OF THE
COMPUTER SCIENCE DEPARTMENT
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF
DOCTOR OF SCIENCE IN COMPUTER SCIENCE
UNIVERSITY OF MASSACHUSETTS LOWELL
1998

Dissertation Director: Georges G. Grinstein, Ph.D.
Professor, Department of Computer Science

Abstract

Database exploration is the process of extracting knowledge from databases using visual, analytic and database tools. Database exploration has two main modeling components: a *systems integration model* that integrates the various tools in support of data exploration tasks, and a *user-centered model* of the data exploration process itself. This thesis takes a dual approach to database exploration by (1) developing and implementing a *systems integration model* based on data exploration tasks, and (2) defining a *generalized data exploration process* model that captures the essence of data exploration sessions. The systems integration model maps between the data models and interaction styles of database and visualization *exploration domains*. The process model describes data exploration without regard for the exploration domain. Two key components of the process model are a set of exploration domain-independent *metrics* that characterize elements of the process, and a set of data exploration *interaction patterns* that characterize database exploration sessions. The generalized process model is then applied to the systems model and implementation.

Acknowledgments

There are many people who assisted me in countless ways during the course of this research. First and foremost, I wish to thank my wife, Amy, for her loving support and patience over the many years it took to produce this work. This thesis is dedicated to her. I am equally thankful for our son, Maxwell Edward Lee, our “aggressive explorer” who has put everything into perspective.

I wish to thank my thesis advisor, Georges Grinstein, who has always been encouraging and supportive of this work, and who has always inspired me to think big. I still marvel at the breadth and depth of his intellect, creativity and compassion. I also wish to thank my thesis committee: Dan Bergeron, Haim Levkowitz and John Sieg, for their guidance in crafting this thesis. I consider myself fortunate to be able to work with such outstanding teachers.

I wish to thank the NASA Graduate Student Researchers Program for supporting three years of this effort, the financial assistance from the University of Massachusetts Lowell, and the support of past and present employers: Paul Breen (MITRE), Ed Campbell (BBN) and Lorne Grant (Spacetec IMC).

I wish to thank my sources of inspiration: Steven Munno, Wassily Kandinsky, Alexander Calder, George Chaikin, Michael Leyton, John Becker, Ernest Wantuch, Edward Abbey, SUWA and the Colorado Plateau. Finally, I wish to thank David Salzman for lighting the fire...

Table of Contents

	<u>Page</u>
List of Tables	vi
List of Figures.....	vii
CHAPTER 1 - INTRODUCTION.....	1
1.1 Problem Statement and Motivation.....	1
1.2 Thesis Goals.....	4
1.3 Thesis Contributions	5
1.4 Thesis Overview.....	7
CHAPTER 2 - LITERATURE REVIEW	10
2.1 Definition of Database Exploration	10
2.2 User-Centric Aspects of Database Exploration	14
2.3 Data-Centric Aspects of Database Exploration.....	39
2.4 Literature Review Summary.....	58
CHAPTER 3 - A SYSTEMS MODEL FOR DATABASE EXPLORATION.....	60
3.1 A Data Exploration Scenario	60
3.2 A User-Centered Analysis of System Components.....	63
3.3 Exbase: An Integrated Database Exploration Environment	74
3.4 Systems Model Summary	99
3.5 Towards a Data Exploration Process Model	102
CHAPTER 4 - A GENERALIZED DATA EXPLORATION (GDE) MODEL.....	103
4.1 Data and Metadata.....	103
4.2 Data Entities	109
4.3 Data Derivations and Sequences.....	112
4.4 Data Derivation Graphs.....	120

	<u>Page</u>
4.5 Forward Derivation Paths.....	131
4.6 GDE Model Summary.....	136
CHAPTER 5 - DESCRIBING DATA EXPLORATION SESSIONS	138
5.1 Data Exploration State Revisited.....	138
5.2 Representing Data Exploration Sessions	140
5.3 Data Exploration Scope.....	146
5.4 Data Exploration Metrics.....	149
5.5 A Data Exploration Calculus	166
5.6 Data Continuity.....	174
5.7 Chapter Summary.....	179
CHAPTER 6 - APPLYING THE GDE MODEL AND METRICS.....	181
6.1 The General Approach.....	181
6.2 The Relational Database Specialization.....	191
6.3 The Iconographic Visualization Specialization	198
6.4 Interaction Patterns in Database Exploration.....	204
6.4 Chapter Summary.....	215
CHAPTER 7 - CONCLUSION	216
7.1 Thesis Summary	216
7.2 The Research in Context.....	223
7.3 Thesis Contributions	225
7.4 Future Work.....	227
7.5 Concluding Thoughts	233
Literature Cited	234
Additional Literature Used but not Cited	244
Biographical Sketch of the Author.....	245

List of Tables

	<u>Page</u>
Table 2-1. Comparison of task-level user-data interactions.....	21
Table 2-2. Comparison of approaches to process-level database interactions.....	36
Table 5-1. Visual representation taxonomy of forward derivation component graph.	145
Table 5-2. Vertex-based metrics summary.	153
Table 5-3. The derivation directional unit vectors of Figure 5-14.	172
Table 6-1. Metric summary for the graph of Figure 6-1.....	186

List of Figures

	<u>Page</u>
Figure 2-1. User-Data Interaction Through Visualization and Database Systems.	13
Figure 2-2. Data exploration interaction framework.	15
Figure 2-3. Springmeyer’s task taxonomy (from Springmeyer 1992).	17
Figure 2-4. The stages of user activity in the performance of a task.	22
Figure 2-5. Components of a <i>problem behavior graph</i>	24
Figure 2-6. The problem behavior graph for the cryptarithmic problem	25
Figure 2-7. The visualization pipeline and cycle models.	26
Figure 2-8 Scientific visualization process network of Felger and Astheimer (1991).	28
Figure 2-9. The history tree process model of Brodlie et al. (1993).	28
Figure 2-10. Scanning search strategy of Canter et al. (1985).	29
Figure 2-11. The data derivation model of Hachem et al. (1994).	30
Figure 2-12. Browsing session model of Kersten and de Boer (1994).	31
Figure 2-13. The history design thread of Chiueh and Katz (1994).	32
Figure 2-14. A data state tree from Carr et al. (1986).	33
Figure 2-15. An AnalysisMap, adapted from Oldford and Peters (1986).	34
Figure 2-16. GuideMaps and WorkMaps from Young and Lubinsky (1995).	35
Figure 2-17. The impedance mismatch between database and visualization systems.	41
Figure 2-18. KDD system autonomy versus versatility, from Matheus et al. (1993).	49
Figure 3-1. The mapping between database relation and a line-oriented icon.	68
Figure 3-2. An Exvis visualization of two MRI scans.	69
Figure 3-3. The Exbase Interface, with primary data objects and operations.	75
Figure 3-4. Exbase view structures, relevant transformations and interaction paths.	76
Figure 3-5. The components of the Exbase local database view.	82
Figure 3-6. Exbase Visualization View hierarchy.	89
Figure 3-7. The Exbase Database Query user interface.	91
Figure 3-8. The Exbase Visualization Manager user interface.	92
Figure 3-9. The Exbase Visualization View user interface, with query visualization.	93
Figure 3-10. Zooming in on coordinate axes-mapped attributes and one other attribute.	94
Figure 3-11. Zooming in on all data attributes.	95
Figure 3-12. Changing the data-to-visual primitive mapping.	96
Figure 3-13. Increasing the <i>icon radius</i> graphical display setting.	97
Figure 3-14 (a). The Exbase Session Log.	99
Figure 3-14(b). The Exbase Session Log (continued).	100
Figure 4-1. Various kinds of d-derivation mappings.	113
Figure 4-2. A graphical representation of a d-derivation sequence.	116
Figure 4-3. Graphical representation of a d-graph.	120
Figure 4-4. Creating a congruence graph.	122
Figure 4-5. Creating a similarity graph.	123

Figure 4-6. A simple forward derivation component graph.....	127
Figure 4-7. Graphical representations of a path and two semipaths.....	129
Figure 4-8. A forward derivation path between two sets of data entities.....	131
Figure 4-9. Completing a forward derivation.....	133
Figure 4-10. The data lineage graph corresponding to Figure 4-6.....	135
Figure 5-1. Visualization of the forward derivation component graph of Figure 4-4. ...	141
Figure 5-2. Bounds on repositioning vertices with increasing branching.....	143
Figure 5-3. The graph of Figure 5-1 visualized as a summary graph.....	144
Figure 5-4. A possible compressed summary graph of Figure 5-3.....	145
Figure 5-5. The local scope of a single target vertex.....	147
Figure 5-6. The derivation path scope of a target forward derivation path.....	148
Figure 5-7. A difference between indegree and number of incident derivations.....	151
Figure 5-8. The fundamental vertex-based metrics and their relationship.....	153
Figure 5-9. Comparing divergent vertices with the Derivation Ratio.....	156
Figure 5-10. Comparing convergent vertices with the Derivation Ratio.....	157
Figure 5-11. Determining the derivation path depth.....	160
Figure 5-12. A one-dimensional view of data exploration.....	168
Figure 5-13. One-dimensional derivation speed and acceleration examples.....	170
Figure 5-14. n-dimensional derivation directional unit vector example.....	172
Figure 5-15. Examples of data continuity.....	176
Figure 5-16. Temporal data continuity examples.....	178
Figure 6-1. Data exploration session summary graph example.....	184
Figure 6-2. Applying the data exploration calculus.....	189
Figure 6-3. Relational database specialization labeling scheme.....	192
Figure 6-4. The <i>zoom in</i> interaction pattern.....	194
Figure 6-5. The <i>zoom out</i> relational database interaction pattern.....	194
Figure 6-6. Relational database data panning interaction patterns.....	196
Figure 6-7. Iconographic visualization specification labeling scheme.....	199
Figure 6-8. The spatial <i>zoom in</i> operation.....	200
Figure 6-9. Spatial <i>zoom out</i> operation.....	201
Figure 6-10. Spatial panning operation.....	202
Figure 6-11. The fundamental database-visualization data exploration.....	204
Figure 6-12. Canonical database exploration session graph example.....	206
Figure 6-13. Refinement before exploration pattern.....	207
Figure 6-14. Querying a visualization derivation edge.....	209
Figure 6-15. Canonical Exbase exploration session graph example.....	211
Figure 6-16. Dynamic query interaction pattern example.....	213

Figure 2-5: from Human Problem Solving, copyright 1972 Prentice Hall Publishers, reproduced by permission of Prentice-Hall Inc., Upper Saddle River, NJ 07458.