

# **Interactive Visual Data Query & Exploration**

*Techniques for visual data analytics through visual query modelling and multidimensional data interaction*

**Phi Giang Pham**

Supervisor: Associate Professor Dr. **Mao Lin Huang**

A thesis submitted in partial fulfilment of the requirements for the degree of

Doctor of Philosophy

in

the Faculty of Engineering and Information Technology

**University of Technology Sydney**

Sydney, Australia 2018

# Certificate of Original Authorship

I certify that the work in this thesis has not previously been submitted for a degree nor has it been submitted as part of requirements for a degree except as fully acknowledged within the text.

I also certify that the thesis has been written by me. Any help that I have received in my research work and the preparation of the thesis itself has been acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis.

Signature of candidate:

Production Note:

Signature removed prior to publication.

Phi Giang Pham

Date: 22 – Jan – 2018

# Acknowledgement

Today is, for me, the day of a beautiful memory which would be unforgettable during my life time. This is because I am here and writing the last but not least of the significant parts of my dissertation that is about the acknowledgement expression for the completing stage of my interesting Ph.D. study. Four years ago, I had not believed and imagined what I could and have reached as today until there was a person who appeared and changed my mind.

Absolutely, the man with the role of my supervisor is Associate Professor Mao Lin Huang, to whom I would like to express my genuine gratefulness firstly. Thanks to his advanced academic guidance, mental encouragement, especially free and active working style deployment, I have learned and experienced plenty of self-study and research methodologies and optimized the strength of mine in order to overcome the research challenges and reach the excellent achievement of today.

Additionally, I would like to thank all of my colleges who greatly supported me during the candidate in the sharing of knowledge, solving the technical problems and dealing with the life issues. I also would like to thank all of the staffs who are working in the school of Software, FEIT, UTS for their help in the administrative and financial procedure.

Finally, it is unexplainable by words and languages actually that I would like to thank all of my family members, who are always beside me, look after me, and love me, especially the meaningful accompany of my wife Le Thu Trang Ho and my daughter Mai Thanh Pham. Without all of them, my Ph.D. study could not be started and completed successfully.

Thanks for all.

# Table of Contents

<b>List of Algorithms</b> .....	viii
<b>List of Figures</b> .....	ix
<b>List of Tables</b> .....	xv
<b>Abstract</b> .....	xvi
<b>Chapter 1 Introduction</b> .....	<b>1</b>
1.1 From Information Visualization to Visual Queries.....	1
1.1.1 Information Visualization and Scientific Visualization.....	2
1.1.2 Visual Queries .....	5
1.2 Problem Statement.....	6
1.3 Challenges and Goals.....	7
1.4 Contributions.....	8
1.5 Skeleton.....	10
<b>Chapter 2 Background</b> .....	<b>12</b>
2.1 Terminology Definitions.....	12
2.2 Relational Data Visualization .....	13
2.2.1 Relational Data Model Visualization.....	13
2.2.2 Relational Data Mapping .....	14
2.2.3 Relational Data Cleaning Processes .....	16
2.2.4 Discussion.....	17
2.3 Multiple Dimensional Visualization .....	17
2.3.1 Parallel Coordinates.....	18
2.3.2 Scatterplots and Scatterplot Matrices .....	19
2.3.3 Star Coordinates.....	21
2.3.4 TableLens .....	22
2.3.5 Discussion.....	23
<b>Chapter 3 A Framework of Visual Data Exploration</b> .....	<b>24</b>
3.1 Introduction.....	24

3.2 A Multiple-Visual-Context Framework of Data Exploration .....	26
3.2.1 Data-Model Context .....	26
3.2.2 Multiple-Dimension Context .....	27
3.2.3 Pairwise-Dimension Context .....	27
3.3 Discussion .....	27
<b>Chapter 4 A New Interactive Visual Query for Relational Data Models .....</b>	<b>29</b>
4.1 Revisiting Visual Queries for Relational Data .....	29
4.2 Data Model Visualization .....	34
4.2.1 Coordinating Context Views .....	34
4.2.2 Node-Link Graph Design for Relational Data Models .....	36
4.2.3 Data Mapping for Visualization .....	42
4.2.4 Discussion .....	47
4.3 Query Interaction .....	47
4.3.1 Interaction Model for Coordinating Context Views .....	49
4.3.2 Interaction Model for Node-Link-Based Queries .....	50
4.3.3 Incremental Data Exploration .....	56
4.3.4 Discussion .....	57
4.4 Visual Navigation Methods .....	58
4.4.1 Focus + Context .....	58
4.4.2 Zooming and Filtering .....	59
4.5 Query Implementation .....	60
4.5.1 A System Framework for Visual Queries with Coordinating Contexts ...	60
4.6 Summary .....	62
<b>Chapter 5 New Interactive Visual Queries for Multi-Dimensional Data.....</b>	<b>63</b>
5.1 Revisiting Manipulation on Multi-Dimensional Data .....	64
5.1.1 Parallel coordinate Interaction .....	64
5.1.2 Scatterplot Interaction .....	67
5.2 Quantitative Visualization with SumUp .....	69

5.2.1 Double Layer Views .....	71
5.2.2 Parallel Coordinate and Stacked Bar Integration.....	72
5.2.3 Data Mapping for Visualization .....	73
5.2.4 Discussion.....	75
5.3 Query Interaction on Parallel Coordinates with Quantitative Approach .....	75
5.3.1 Quantitative Visual Queries by Brushing .....	77
5.3.2 Discussion.....	81
5.4 Query Interaction on Scatterplots by FigAxis.....	82
5.4.1 Quantitative Visual Queries by Zooming .....	82
5.4.2 Data Mapping .....	88
5.5 Visual Enhancement Methods .....	88
5.5.1 Parallel Coordinate Scalability .....	88
5.5.2 Scatterplot Navigation .....	91
5.6 Query Implementation .....	94
5.6.1 A System Framework for Visual Queries with Quantitative Approach ...	94
5.7 Summary .....	96
<b>Chapter 6 Case Studies.....</b>	<b>97</b>
6.1 Case Study 1: Visual Data Exploration with Relational Models .....	97
6.1.1 Scenario 1 .....	97
6.1.2 Scenario 2 .....	99
6.1.3 Scenario 3 .....	101
6.1.4 Scenario 4 .....	104
6.1.5 Discussion.....	105
6.2 Case Study 2: Visual Analysis of Multiple-Dimensional Data .....	106
6.2.1 Multiple-Attribute Comparison .....	106
6.2.2 Correlative Analysis .....	108
6.2.3 Flexible Data Support .....	109
6.2.4 Discussion.....	110

6.3 Case Study 3: Interactive Data Exploration of Multiple Visual Contexts .....	110
6.3.1 In-depth Exploration on Multiple Dimensions and Pairwise Comparison	111
6.3.2 Multiple-Context Queries of Data Models and Multi-Dimensional Data	115
6.3.3 Discussion.....	120
6.4 Summary .....	121
<b>Chapter 7 Evaluations .....</b>	<b>122</b>
7.1 Space-Efficient Visualization .....	122
7.1.1 Dynamic Representation.....	122
7.1.2 Layering Display + Sharing Axes.....	125
7.2 Distinctive Features .....	127
7.2.1 Relational Query Making through Node-Link Graphics .....	127
7.2.2 Quantitative Query Making through Parallel Coordinates .....	128
7.2.3 Quantitative Query Making through Scatterplots.....	129
7.3 Friendliness of Techniques .....	129
7.3.1 Query Making with Node-Link Graphics.....	129
7.3.2 Query Making with Parallel Coordinates .....	132
7.3.3 Query Making with Scatterplots.....	134
7.4 Discussion.....	136
<b>Chapter 8 Extended work .....</b>	<b>137</b>
8.1 Introduction.....	137
8.2 Review of Tag Visualization.....	138
8.3 Ranking Visualization Technique .....	139
8.3.1 Basic Design and Interaction .....	139
8.3.2 Grouped-Score Rankings.....	141
8.4 Visual Enhancement .....	142
8.5 Data Mapping.....	144
8.6 Case Study .....	144

8.6.1 Overall Contribution Rankings .....	144
8.6.2 Topic Contribution Rankings .....	147
8.7 Evaluation .....	148
8.7.1 Study Setup and Procedure .....	149
8.7.2 Result .....	149
8.8 Discussion .....	150
<b>Chapter 9 Conclusion .....</b>	<b>151</b>
9.1 Summary .....	151
9.2 Final Conclusion .....	152
<b>List of Publications .....</b>	<b>154</b>
<b>Bibliography .....</b>	<b>155</b>



# List of Algorithms

Algorithm 4.1 The algorithm of setting relationships from the data schema information .....	43
Algorithm 4.2 The algorithm of checking a foreign key from the data schema information .....	44
Algorithm 4.3 The algorithm of setting nodes from the data schema information .....	45
Algorithm 4.4 The algorithm of the visualization mapping from the data of a query result .....	47
Algorithm 5.1 The procedure of recording the interaction on the parallel coordinate dimensions .....	80
Algorithm 5.2 The procedure of displaying the query results for the stacked bar interaction .....	81
Algorithm 5.3 The procedure of recording the interaction on FigAxis .....	87
Algorithm 6.1 Scenario completion procedures of MCquery and a data query tool .....	105

# List of Figures

Figure 1.1 The visualization of car crashing experiment .....	2
Figure 1.2 A tree map sample .....	3
Figure 1.3 The focus layer display in a biological study .....	4
Figure 1.4 The icon-based representations for a control query .....	6
Figure 2.1 A sample of the ER application .....	13
Figure 2.2 An entire employment of a simple ER diagram .....	14
Figure 2.3 The impact of the attributed visualization on node-link graphs ..	15
Figure 2.4 The relational context visualization for entity resolution .....	16
Figure 2.5 The visual calculation of the y-coordinate of a basic data point (di) in parallel coordinates .....	18
Figure 2.6 An application of parallel coordinates with nine dimensions and around four hundred instances .....	19
Figure 2.7 The scatterplots with different shapes of data points.....	20
Figure 2.8 The scatterplot matrix with 4 dimensions.....	21
Figure 2.9 The star coordinates with 8 dimensions.....	22
Figure 2.10 A sample layout of the TableLens employment .....	23
Figure 3.1 The proposed multiple-visual-context framework of relational data exploration.....	26
Figure 4.1 The relational data query interface of MS Access 2010.....	30
Figure 4.2 The relational data query interface of QGraph .....	31
Figure 4.3 The relational data query interface of Dataplay.....	31

Figure 4.4 The relational data query interface based on HV.....	32
Figure 4.5 The relational data query interface based on a node-link graph ..	33
Figure 4.6 The main user interface for the coordinating visual contexts of data models and query results by technique MCquery. ....	35
Figure 4.7 The relational schema of six tables Payments, Customers, Countries, Orders, OrderDetails, and Products used in the samples of this chapter. ....	37
Figure 4.8 A sample of the data model representation for six tables. ....	38
Figure 4.9 A sample of the result graph corresponding to three tables countries (a blue node), customers (an orange node), and products (the green nodes) .....	39
Figure 4.10 A query formulation model.....	48
Figure 4.11 The query interaction model proposed for the query module of Microsoft Access and MySQL .....	49
Figure 4.12 The new model for the data exploration by interaction on the coordinating visual contexts of data models and query results. ....	49
Figure 4.13 The instance of a query formulation with the finding component in the data model context .....	51
Figure 4.14 The instance of a query formulation with the condition component in the data model context.....	52
Figure 4.15 The instance of removing a link from a query .....	53
Figure 4.16 The filtering feature in the data model context.....	54
Figure 4.17 The instance of query interaction in the query result context....	55
Figure 4.18 The logical-frame-based exploration of a huge graph .....	56

Figure 4.19 The proposed system framework for visual queries with the coordinating context views.....	60
Figure 5.1 The direct manipulation with rectangle drawing .....	64
Figure 5.2 The density based filtering in parallel coordinates .....	65
Figure 5.3 A focus+context visualization model in parallel coordinates .....	66
Figure 5.4 A dimensional tree of visual hierarchical dimensional reduction	66
Figure 5.5 The scatterplots with dynamic queries.....	67
Figure 5.6 The scatterplots with data point selection optimization.....	68
Figure 5.7 The scatterplots with display space transformation .....	68
Figure 5.8 The bar and stacked bar layout with various arrangements for visual rankings.....	70
Figure 5.9 The TableLens layout .....	70
Figure 5.10 The main layout design of the SumUp user interface.....	71
Figure 5.11 An instance of the SumUp query comparing the car models of three representatives Toyota of Japan (green), Volkswagen of Europe (orange), and Ford of USA (red) .....	73
Figure 5.12 The data structure for considered polyline ranges .....	74
Figure 5.13 The data matrix for stacked bar representation.....	74
Figure 5.14 The parallel coordinates with box-plot embedded for data instance summary .....	77
Figure 5.15 The interaction model of query interaction on the double layer views.....	79
Figure 5.16 The linked views of scatterplots with various graphs.....	82

Figure 5.17 The procedure of multiple-view matching between scatterplots and other graphics .....	83
Figure 5.18 The overview of the FigAxis layout design .....	84
Figure 5.19 The layout of a FigAxis application for the correlative comparison of new car model delivery in term of Horsepower and Weight with the targets of original brands from USA, Europe, and Japan.....	85
Figure 5.20 The zooming level measurement background .....	86
Figure 5.21 The filtering feature of SumUp applied in the visual analysis of Census income data concerning Income, Hourperweek, Age, and Sex towards Occupation .....	90
Figure 5.22 The zooming and panning feature of FigAxis in the proximity navigation .....	91
Figure 5.23 The colour-based highlight of the car models delivered by USA (the green plotted points and the green stacked bars). .....	92
Figure 5.24 The extended FigAxis layout of the visual comparison of the car model delivery in term of Year and Cylinder .....	93
Figure 5.25 The system framework proposed for the visual query deployment of SumUp .....	95
Figure 6.1 The query interaction on the relationship representations of store, staff, city, and country .....	98
Figure 6.2 The flexible interaction on the coordinating visual contexts of the data model and query results in Scenario 2 .....	100
Figure 6.3 The filtering feature applied for countries China and India.....	101
Figure 6.4 The query interaction with the highlighted dimensions of film, category, and actor .....	103

Figure 6.5 The relationship recognition in the query result of categories Children and Comedy.....	104
Figure 6.6 The multiple attribute comparison for the number of models of six Japanese brands based on Cylinder, Horsepower, Weight, and Year .....	107
Figure 6.7 The correlative analyses of Weight and MPG, Weight and Horsepower, and Weight and Displacement. ....	108
Figure 6.8 The data summary with flexible data support to explore Income and Workclass towards the ages of the population in United States.....	109
Figure 6.9 The statistical parallel coordinates of the United States Census income data towards Sex and Education.....	110
Figure 6.10 The leve-1 summary of Age and 40-and-over Hoursperweek.	112
Figure 6.11 The leve-6 summary of Age and 40-and-over Hoursperweek.	113
Figure 6.12 The quantitative plotting comparison between Occupation and 40-50 HoursPerWeek .....	114
Figure 6.13 The data-model context of customer, film, category, actor, and store .....	115
Figure 6.14 the multi-dimension context of categoryname, filmreleaseyear, and storeaddress .....	116
Figure 6.15 The pairwise comparison of categoryname and filmreplacementcost .....	118
Figure 6.16 The name of the impacted films of Drama and Family in the comparison of costs 10.99 and 24.99 visualized in the data-model context of MCquery.....	120
Figure 7.1 The space-saving rates in term of the link display of MCquery	124
Figure 7.2 The space-saving rates in term of the node display of MCquery	125

Figure 7.3 The detailed comparison between the IT and BA groups on the MCquery feedbacks.....	130
Figure 7.4 The task completion time of the MCquery usage for the IT and BA groups .....	131
Figure 7.5 The feedback result of the friendliness comparison between the FigAxis layout and the multi-view layout.....	135
Figure 8.1 The basic layout design of Qstack .....	140
Figure 8.2 An instance of the grouped-score visual ranking.....	141
Figure 8.3 A stacked bar chart with multi-tag filtering.....	143
Figure 8.4 The scaling components of Qstack .....	143
Figure 8.5 The overall contribution ranking for price A.....	145
Figure 8.6 The overall contribution ranking for price B .....	146
Figure 8.7 The sorting view by tag park (red bars) and tag flower (green bars) with multi-baselines and flexible scales.....	148
Figure 8.8 The brushing across categories 4, 5, and 6 .....	148

# List of Tables

Table 4.1 The data of the query result visualized in Figure 4.9. ....	40
Table 4.2 The 10-categorical colour scheme of d3js library .....	41
Table 4.3 The 20-categorical colour scheme of d3js library .....	42
Table 5.1 The FigAxis data support summary .....	88
Table 7.1 The parameter values chosen for the space-saving assessment in term of the node display of MCquery .....	124
Table 7.2. The feature comparison of MCquery, the visual query tool of Microsoft Access, and Ploceus.....	127
Table 7.3 The feature comparison of SumUp, the tool of Siirtola (2002), and Ho et al. (2011).....	128



# Abstract

The direct data manipulation through visualization and associated navigation techniques has been implemented for many years. However, these methods are not uniformly discussed in the context of user interface design. During the history of user interface development, the interaction between humans and computers is almost to be done through software widgets. Since in the last decade, many advanced data visualization and interaction techniques have been developed, now it is the time to bring them into the formal discussion about the context of user interface design, data queries, and data manipulation. The dissertation attempts to fulfill the gap between visual user interface design and interactive data visualization.

In relational data queries, many visualization techniques have featured advanced interactive operation; however, a majority of those would concentrate on the traditional style, instead of a modern approach. This is the reason why today in visual analytics truly direct manipulation is highly encouraged, instead of the conventional methods.

This dissertation focuses on the investigation of modern data query approaches. It attempts to model the new data query methods that apply those advanced visualization and interaction techniques to facilitate the data analysis procedures. The second contribution of the dissertation is the design of new interaction methods for multi-dimensional data visualization.

We first introduce a new framework which includes straightforward manipulation techniques for relational data discovery. These novel techniques, named *MCquery*, *SumUp*, and *FigAxis*, are exclusively developed for the key characteristics of relational data such as *data models* and *data dimensions*. The core methodology is about interactive visual query design based upon node-link graphics, parallel coordinate geometries, and scatterplot visualization, where the direct interaction is performed by friendly action such as clicks and brushes. The tools materialized from these techniques can help to reduce users' cognitive and behavioral effort efficiently in dealing with the issues of information search-retrieval, quantitative data analysis, and correlation examination.