Doctoral Dissertation

Curran Kelleher

4/9/2014

Abstract

There is immense potential value in data that is not being realized. While many data sets are available, it is difficult to realize their full value because they are made available using many different formats, protocols and vocabularies. The heterogeneity of formats, protocols and vocabularies makes it difficult to combine data sets together and hinders the development of data visualization software. While it is straightforward to produce static visualizations of just about any data set by customizing existing examples, there is a lack of generalized visualization software that supports the creation of interactive visualizations and visualization dashboards with multiple linked views. The contribution of this dissertation is a collection of data structures and algorithms supporting integration and interactive visualization of many data sets using interactive visualization dashboards with multiple linked views. A proof of concept implementation demonstrates support for several public data sets and well known visualization techniques.

Contents

1 Introduction

The contributions of this dissertation are novel data structures and algorithms for integration and interactive visualization of many data sets from multiple sources, based on the data cube concept. The proposed data representation framework will allow data sets to be combined together and visualized using interactive visualization dashboards like the one described above, giving users the sense that the data exists within a single unified structure. The framework is designed to be able to represent and integrate an arbitrary number of data sets created independently of one another, and expose the integrated structure to reusable visualization tools that can be combined together in dashboard layouts with multiple linked views using existing interaction techniques such as brushing and linking. The proposed data representation and visualization framework is fundamentally new, and will allow heterogeneous data sets to be explored in a unified way that was never before possible.

The overall goal of this work is to build digital telescope into the universe of phenomena on Earth via publicly available data. For example, consider public data sources such as the United Nations, the US Census, the US Bureau of Labor Statistics, or the US Centers for Disease Control. These organizations and hundreds of others around the world provide publicly available data about various topics including population statistics, public health, distribution of wealth, quality of life, economics, the environment, and many others. By unifying these data sources and providing users with tools to explore and present the data visually, a deeper understanding of the world can be gleaned through the lens of public data. The focus of this dissertation is on applications involving public data, however the techniques introduced can be applied to any data sets that can be conceptually modeled as data cubes, regardless of whether they are public or private.

Consider the data from the US Census that covers population statistics for US States from 1950 to 2010. Consider also population statistics from the United Nations covering World Countries from 1970 to 2012. These two data sets may use different identifiers for years and geographic regions, but they cover an overlapping conceptual data space of time, geography and population. From these two data sets it is possible to create a visualization dashboard with a map of the world showing population as color and a corresponding line graph showing population for each region as lines. If the user views the whole world, the UN population data is shown for each country. If the user zooms into the US, US Census data is shown for each state. If the user selects a point of time in the line graph, the data shown on the map is from that point in time. If the user pans and zooms on the map, the lines in the line graph update to only show the regions visible on the map. This is one example of an interactive visualization dashboard with multiple linked views (the timeline and map views) operating over multiple data sets integrated from different sources (the United Nations population data and the US Census population data).

TODO implement this and put screenshot image here

Data cubes, also known as OLAP (OnLine Analytical Processing) cubes, can represent data that contains measures aggregated (typically using sum or average) along categorical hierarchies. The data cube concept emerged from the field of data warehousing as a way to summarize transactional data, allowing analysts to get a bird's eye view of company activities. The term OLAP stands in contrast to the term OLTP (OnLine Transaction Processing), which is the part of the data warehouse system that ingests and stores data at the level of individual transactions or events. After the ETL (Extract, Transform and Load) phase of the data warehouse flow, the data is analyzed by computing a data cube from the transactional data.

The data cube concept and structure can be used to model existing data sets as well. Publicly available data sets (often termed "statistical data") may be considered as pre-computed data cubes if they contain aggregated measures (also called "indicators", "metrics" or "statistics") across time, geographic space or other dimensions such as gender, age range, ethnicity or industry sector. Any categorization scheme containing distinct entities, organized as an unorered collection, an ordered collection, or a hierarchy can be modeled as a dimension. Any numeric value that represents an aggregated statistical summary using sum, average, or other aggregation operator can be modeled as a measure.

With this approach, it is possible to model many data sets together using shared dimensions and measures. This will allow integration of many data sets together in a single unified structure. Existing data cube technologies assume that data cubes will be computed from a relational source, and are not designed to handle integration of pre-computed data cubes that may use inconsistent identifiers for common dimensions and inconsistent scaling factors for common measures. Therefore the application of the data cube concept to integration and visualization of many pre-computed data cubes, while theoretically plausible, requires the development of novel data structures and algorithms that extend the data cube model to handle integration of pre-computed data cubes that may use inconsistent identifiers for common dimensions and inconsistent scaling factors for common measures.

The data cube structure lends itself particularly well to visualization. Long standing perception-based data visualization theory presented by Bertin [2] and Mackinlay [13] identify effective ways to visually encode data based on data fields that are nominal, ordinal or quantitative. Visualization techniques have been explored for hierarchical (tree-based) data as well [8]. Data cubes can contain data of these types. Therefore existing visualization theory can be applied to the data cube model to determine which visualizations are appropriate for representing which data, depending on its data cube structure. This topic is discussed in section 4.

2 Related Work

Previous work relating to this dissertation falls into four major categories:

- data representation - structures, models and formats for data
- data integration - merging data from many sources
- data visualization - transforming data into interactive graphics
- Web graphics technology - HTML5 graphics APIs and libraries

2.1 Data Representation

Relational database systems provide a mature data management solution and are widely adopted [14]. The relational model has well understood theoretical underpinnings such as relational algebra [5]. Data warehouse systems are typically built on the relational model and are augmented by multi-scale aggregated data structures called data cubes, also known as OLAP (OnLine Analytical Processing) cubes [9, 6]. Data cubes contain summaries of the collection of facts stored in a relational database [4]. For example, a data cube may contain how much profit was made from month to month subdivided by product category, while the relational database may contain the information associated with each individual transaction.

Because data cubes provide an abstraction that handles aggregation, they are a widely used method of data abstraction for supporting visualization and analysis tasks [15]. Kimball pioneered the area of “Dimensional Modeling”, which concerns constructing data warehouse schemas amenable to data cube construction and analysis [11]. Data cubes have been implemented in a variety of different systems, so effort has been made to discover unified conceptual or mathematical models that can characterize many implementations [7, 17, 16, 12, 1, 10, 3]. The data cube structure has also been used to model user Web browsing sessions to support data mining algorithms for Web prefetching [18].

2.2 Data Integration

2.3 Data Visualization

2.4 Web Graphics Technology

TODO pull all related work from proposal

3 Data Cube Representation and Integration

3.1 Defining Data Cubes

TODO describe CSV + JSON representation

3.2 Building a Data Cube Index

TODO pseudocode for building the index

3.3 Querying a Data Cube

TODO pseudocode for querying the index

3.4 Integration of Dimensions

TODO include modified pseudocode for UDC index building & query that resolves identifiers

4 Data Cube Visualization

4.1 Relating Data Cubes and Visualization Theory

Data cubes can contain the following kinds of data:

- *nominal* - unordered collections of categories
- *ordinal* - ordered collections of categories
- *hierarchical* - collections of categories organized as trees
- *quantitative* - continuously varying numeric values

Data cube dimensions can be nominal, ordinal or hierarchical. Data cube measures are always quantitative. This mapping relates data cubes to data types that have been well studied in the literature on visualization theory [2, 13, 8].

TODO add table with (nominal, ... , quantitative) vs. (color, value, position, size, connection, containment, ...)

4.2 A Visualization Taxonomy by Data Cube Structure

TODO add table with (Visualization, Data Cube Structure) referencing visualizations in other sections

5 Visualization Dashboard Infrastructure

5.1 Model Driven Visualizations

TODO include model driven bar chart pseudocode

5.2 Functional Reactive Visualizations

TODO include data flow graph for bar chart **TODO** Discuss the ”when” Functional Reactive Operator **TODO** Cite functional reactive animation paper **TODO** include bar chart pseudocode using ”when”

5.3 Dashboard Layout using Nested Boxes

TODO Include nested box layout pseudocode

5.4 Multiple Linked Views

TODO Include generic linking pseudocode using ”when”

5.5 Dynamic Dashboard Configuration

TODO Include pseudocode for computing configuration diffs

6 Data Sets

6.1 United Nations Population Estimates

6.2 United Nations Millenium Development Goals

6.3 United States Census Population Estimates

TODO include figures/usCensusPopulationByState.png **TODO** import data from http://www.census.gov/popest/data/state/totals/2013/index.html

6.4 US Central Intelligence Agency World Factbook

TODO import subsets of the data

6.5 US Centers for Disease Control Causes of Death

TODO generalize stacked area and tree vis

6.6 W3Schools Browser Market Share

TODO import this data completely

6.7 Natural Earth

TODO include figures/naturalEarth.png **TODO** discuss data transformation process

7 Visualizations

TODO discuss pseudocode conventions **TODO** include pseudocode for each visualization

7.1 Bar Chart

7.2 Scatter Plot

7.3 Line Chart

7.4 Choropleth Map

7.5 Stacked Area Plot

7.6 TreeMap

7.7 Node Link Tree

7.8 Radial Tree

7.9 Icicle Plot

7.10 TreeMap

7.11 Parallel Coordinates

8 Visualization Dashboards

References

[1] Rakesh Agrawal, Ashish Gupta, and Sunita Sarawagi. Modeling multidimensional databases. In *Data Engineering, 1997. Proceedings. 13th International Conference on*, pages 232–243. IEEE, 1997.

[2] Jacques Bertin. Semiology of graphics: diagrams, networks, maps. 1983.

[3] Markus Blaschka, Carsten Sapia, Gabriele Hofling, and Barbara Dinter. Finding your way through multidimensional data models. In *Database and Expert Systems Applications, 1998. Proceedings. Ninth International Workshop on*, pages 198–203. IEEE, 1998.

[4] Surajit Chaudhuri and Umeshwar Dayal. An overview of data warehousing and olap technology. *ACM Sigmod record*, 26(1):65–74, 1997.

[5] James Clifford and Abdullah Uz Tansel. On an algebra for historical relational databases: two views. In *ACM SIGMOD Record*, volume 14, pages 247–265. ACM, 1985.

[6] Edgar F Codd, Sharon B Codd, and Clynch T Salley. Providing olap (on-line analytical processing) to user-analysts: An it mandate. *Codd and Date*, 32, 1993.

[7] Anindya Datta and Helen Thomas. The cube data model: a conceptual model and algebra for on-line analytical processing in data warehouses. *Decision Support Systems*, 27(3):289–301, 1999.

[8] Martin Graham and Jessie Kennedy. A survey of multiple tree visualisation. *Information Visualization*, 9(4):235–252, 2010.

[9] Jim Gray, Surajit Chaudhuri, Adam Bosworth, Andrew Layman, Don Reichart, Murali Venkatrao, Frank Pellow, and Hamid Pirahesh. Data cube: A relational aggregation operator generalizing group-by, cross-tab, and sub-totals. *Data Mining and Knowledge Discovery*, 1(1):29–53, 1997.

[10] Marc Gyssens and Laks VS Lakshmanan. A foundation for multi-dimensional databases. In *VLDB*, volume 97, pages 106–115, 1997.

[11] Ralph Kimball. *The data warehouse lifecycle toolkit: expert methods for designing, developing, and deploying data warehouses*. Wiley. com, 1998.

[12] Chang Li and X Sean Wang. A data model for supporting on-line analytical processing. In *Proceedings of the fifth international conference on Information and knowledge management*, pages 81–88. ACM, 1996.

[13] Jock Mackinlay. Automating the design of graphical presentations of relational information. *ACM Transactions on Graphics (TOG)*, 5(2):110–141, 1986.

[14] Raghu Ramakrishnan and Johannes Gehrke. *Database management systems*. Osborne/McGraw-Hill, 2000.

[15] Chris Stolte, Diane Tang, and Pat Hanrahan. Multiscale visualization using data cubes. *Visualization and Computer Graphics, IEEE Transactions on*, 9(2):176–187, 2003.

[16] Panos Vassiliadis. Modeling multidimensional databases, cubes and cube operations. In *Scientific and Statistical Database Management, 1998. Proceedings. Tenth International Conference on*, pages 53–62. IEEE, 1998.

[17] Panos Vassiliadis and Timos Sellis. A survey of logical models for olap databases. *ACM Sigmod Record*, 28(4):64–69, 1999.

[18] Qiang Yang, Joshua Zhexue Huang, and Michael Ng. A data cube model for prediction-based web prefetching. *Journal of Intelligent Information Systems*, 20(1):11–30, 2003.