Doctoral Dissertation

Curran Kelleher

4/5/2014

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

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).

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 focus will be on public data, however the techniques can be applied to any data collection that can be conceptually modeled as data cubes. 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 (nested box 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.

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

[1] and Mackinlay [2] identify effective ways to visually encode data based on the various kinds of data fields:

- *nominal* - unordered collections of categories
- *ordinal* - ordered collections of categories
- *quantitative* - continuously varying numeric values

Place	Time	Population	Source URL
India	1950	369880000	www.geohive.com/earth/population3.aspx
China	1950	563000000	geography.about.com/od/populationgeography/a/chinapopulation.htm
USA	1950	150697361	en.wikipedia.org/wiki/1950_United_States_Census
India	2010	1150000000	www.indiaonlinepages.com/population/india-population.html
China	2010	1339724852	en.wikipedia.org/wiki/Demographics_of_China
USA	2010	308745538	en.wikipedia.org/wiki/United_States_Census

References

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

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