T-Cube as an Enabling Technology in Surveillance Applications

Artur Dubrawski, Maheshkumar Sabhnani, Saswati Ray, Josep Roure and Michael Baysek
The Auton Lab, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, 15213

OBJECTIVE

T-Cube, a data structure designed to efficiently represent large collections of temporal data has been shown to benefit surveillance applications involving monitoring sales of over-the-counter medications and emergency department visits [1,2]. In this paper we present efficiencies which can be realized in practical applications of T-Cube beyond its original areas of deployment, and we advocate a widespread use of it as a technology which makes manual ad-hoc lookups as well as many kinds of complex automated analyses feasible.

BACKGROUND

T-Cube is especially useful for rapidly retrieving responses to ad-hoc queries against large datasets of additive time series labeled using a set of categorical attributes. It can be used as a general tool to support any task requiring access to such data. From the application's perspective it is transparent: it acts just like the database itself, but an incredibly quickly responding one. The authors had a chance to put T-Cubes into practical use as an enabling technology in applications requiring massive screening of multidimensional temporal data. These applications include two systems to support monitoring of food and agriculture safety and predictive analytics developed at the US Department of Agriculture [3] and the Food and Drug Administration, as well as a system to monitor and forecast health of a fleet of aircraft operated by the US Air Force [4].

RESULTS

One of those projects involved a data base consisting of datasets each with about 25 demographic attributes of arities varying from 2 to 80, about 12 thousand records of transactions, covering 6 years at a daily resolution. The application called for a massive screening through all combinations of attribute-value pairs of size = 1 and 2, the total number of such combinations exceeding 4.3 million. The involved analytics was based on an expectation-based temporal scan used to detect unusual short-term increases in counts of specific aggregate time series. The total number of individual temporal scan tests for one such data set exceeded 9.3 billion. Each such test involved a Chi-square test of independence performed on a 2-by-2 contingency table formed by the counts corresponding to the time series of interest (one of the 4.3 million series) and the baseline counts, within the current temporal window of interest (one of 2,000+) and outside of it. The complete set of computations, including the time necessary to retrieve and aggregate all the involved time series, compute and store the test results, load source data and build the T-Cube structure, etc., took about 8 hours of time when executed on a dual CPU AMD Opteron 242 1,600 MHz machine, in the 64 bit mode, using 1MB per CPU level 2 cache and 4 GB of main memory, running under Cent OS 4 Linux operating system. If the users chose one of the commercial database tools, the time needed to retrieve the time series data corresponding to one of the involved queries would approach 180 milliseconds. Therefore, without the T-Cube, it would take about 9 days to just to pull all the required time series from the database, not including any processing or execution of statistical tests. That kind of analysis would be considered infeasible without the efficiencies provided by the T-Cube representation.

CONCLUSIONS

T-Cubes are simple to setup and easy to use. Typically, it takes only minutes to build one from data. Database users do not need to define any stored procedures, or materialized views in order to make that happen. Once a T-Cube is built, it is ready to respond to any simple or complex query. The response time speedup has two main benefits: (1) It enables a massive scale statistical mining of large collections of time series data, and (2) It allows the users to perform many complex ad-hoc queries without inconvenient delays. That makes T-Cubes potentially very useful in a range of practical applications, including syndromic surveillance. They have a potential to change the way users (people as well as analytic software systems) deal with the time series datasets.

ACKOWLEDGMENTS

This work was supported by the Centers of Disease Control (award number R01-PH000028), the National Science Foundation (IIS-0325581), the United States Department of Agriculture (533A94031), the United States Air Force (FA8650-05-C-7264) and the Food and Drug Administration (CIO-SP2i-TO-C-2426).

REFERENCES

[1] Sabhnani M., Moore A. and Dubrawski A. Rapid Processing of Ad-hoc Queries against Large Sets of Time Series. Advances in Disease Surveillance 2, 2007.

[2] Sabhnani M., Moore A. and Dubrawski A. T-Cube: A Data Structure for Fast Extraction of Time Series from Large Datasets. Technical Report CMU-ML-07-114, Carnegie Mellon University, 2007.

[3] Roure J., Dubrawski A. and Schneider J. A Study into Detection of Bio-Events in Multiple Streams of Surveillance Data. In D. Zeng et al. (Eds.): BioSurveillance 2007, Lecture Notes in Computer Science 4506, 2007.

[4] Mikus S., Dubrawski A., Sondheimer N., Moyer L, Baysek M., Ostlund J., Stewart T. and Mowry B. Collective Machine Learning for Early Identification of Logistics Crises, Integrated Health Management Conference, Cincinnati, OH, 2007.