Rapid Processing of Ad-Hoc Queries Against Large Sets of Time Series

Maheshkumar R. Sabhnani, Andrew W. Moore, Artur W. Dubrawski

School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, 15213

OBJECTIVE

We present *T-Cube*, a new tool for very fast retrieval and analysis of time series data. Using a novel method of data caching, *T-Cube* performs time series queries approximately 1,000 times faster than standard state-of-the-art data cube technologies. This speedup has two main benefits: it enables fast anomaly detection by simultaneous statistical analysis of many thousands of time series, and it allows public health users to perform many complex, ad hoc time series queries on the fly without inconvenient delays.

BACKGROUND

Time series analysis is very common in syndromic surveillance. Large scale biosurveillance systems typically perform thousands of time series queries per day: for example, monitoring of nationwide over-the-counter (OTC) sales data may require separate time series analyses on tens of thousands of zip codes. More complex query types (e.g. queries over various combinations of patient age, gender, and other characteristics, or spatial scans performed over all potential disease clusters) may require millions of distinct queries. Commercial OLAP databases provide *data cubes* to handle such ad hoc queries [1], but these methods typically suffer from long build times (typically hours), huge memory requirements (requiring the purchase of high-end database servers), and high maintenance costs. Additionally, data cubes typically require 1 second or more to respond to each complex query. This delay is an inconvenience to users who want to perform multiple queries in an online fashion; additionally, data cubes are far too slow for statistical analyses requiring millions of complex queries, which would require days of processing time.

METHOD

All-dimensional (AD)-trees have been successfully used in rapid ad-hoc querying of sets of categorical data in complex machine learning tasks such as Bayesian Network structure learning [2]. We propose using AD-trees as a caching technique for storing time series in *T-Cube*. Report [3] presents technical details of the *T-Cube* design and introduces techniques leading to further performance improvements: specific arrangements of demographic attributes, most-common-value-based pruning, controlling the depth of the tree. They help to balance building time, query response time and physical memory requirements of the tool. The results presented in this paper have been obtained while using *T-Cube* to assist in quickly responding to ad-hoc queries in order to speed up univariate analysis of the aggregated time series in search for unusual patterns.

RESULTS

We investigated the performance of *T-Cube* on three large datasets. The first, emergency patient records from four U.S. states contained 3.4 million records with date (1 year duration at a daily resolution), patient home zipcode (21,179 distinct values), gender (3), age group (3), and syndrome (8). The second, one year nationwide OTC pharmacy sales record had 57 million records with four attributes (date, zipcode, drug category, and promotion flag). Third, we generated a synthetic dataset containing 12 million records with date, zipcode and 29 sparse attributes. Average response time to random complex queries varied between 1 and 5 milliseconds, which was over 1,000 times faster than state-of-the-art commercial data cube packages. Building a *T-Cube* on each of the three data streams took less than 15 minutes, while its memory requirements ranged from 50 to 900MB and they were inversely proportional to the query response time. With *T-Cube*, we could analyze 100,000 distinct time series using complex queries in less than 5 minutes. We have successfully deployed *T-Cubes* as a support tool in follow-up investigations of alerts generated by spatial scan systems [4].

CONCLUSIONS

Rapid aggregation of time series across large data sets is an enabling capability which makes manual lookups as well as many analyses feasible. That is very important in a daily practice of public health monitoring. *T-Cube* can be used to execute analyses of millions of time series in minutes, and it can support real-time manual browsing of data with very little setup time. *T-Cubes* do not require expensive hardware to setup and easy to maintain.

This work was supported by NSF grant IIS-0325581.

REFERENCES

[1] Han J, Kamber M, Data Mining: Concepts and Techniques; Morgan Kaufmann Publishers, 2000.

[2] Moore A, Lee M, Cached sufficient statistics for efficient machine learning with large datasets; Journal of Artificial Intelligence research, 8, 67-91, 1998.

[3] Sabhnani M, Moore A, Dubrawski A, T-Cube: Fast extraction of Time Series from Large Datasets, Technical Report, Carnegie Mellon University, CMU-ML-06-104.

[4] Sabhnani M, Neill D, Moore A, Tsui F, Wagner M, and Espino J, Detecting anomalous patterns in pharmacy retail data; Proceedings of the KDD 2005 Workshop on Data Mining Methods for Anomaly Detection, 2005.