Online Query Processing: A Tutorial

Peter J. Haas
Almaden Research Center
IBM Research Division
peterh@almaden.ibm.com

Joseph M. Hellerstein
Computer Science Division
University of California, Berkeley
jmh@cs.berkeley.edu

1. MOTIVATION

There has been recent interest in techniques for *online query processing*, which produce incrementally refining results for long-running queries. In online *aggregation* queries, incremental results are provided via estimations of the final results, along with statistical confidence intervals for the estimations. In online *enumeration* queries, incremental results can be provided in the form of individual rows or fields of an output relation. In both cases, these results are perceived by the user over time, and hence their improvement in quality should be maximized over time – a goal that is quite different from traditional metrics of system performance. As a corrolary to this interactive feedback, it is natural to provide users of online query processing systems with control over a system while it runs, allowing them to change their performance and query specifications based on the feedback they receive from the system. This research agenda brings together a number of threads in database research, including query processing architectures and algorithms, sampling and statistical estimation, and user interfaces for data visualization and transformation.

The interest in online query processing for data analysis arises naturally both from user behavior and technological pressures. User studies show that data analysis typically proceeds in an iterative fashion, starting with broad "big picture" questions, followed by refined queries chosen based on feedback and domain knowledge, with periodic jumps to start the process over using different information. This kind of interaction favors "quick and dirty" feedback from those queries that are most data-intensive. Data appetites are growing faster than Moore's Law, and hence computing these big queries to completion will remain time-consuming for the foreseeable future. Online query processing is natural in these multi-step data analysis tasks, allowing users to dynamically tradeoff the runtime of each step against the quality of the results seen at that step.

Data analysis is not the only natural application for online query processing. Data cleaning is another large-scale, complex data processing task, which is critical both in data warehousing and federated database design. Data cleaning typically consists of multiple iterations through two time consuming tasks: auditing and mining for discrepancy detection, and transformation specification or view definition. By combining online query processing techniques with clever interfaces, one can "close the loop" between these two tasks: online discrepancy detection algorithms can run over transformed views that are themselves computed online. The result is a unified direct-manipulation experience, in which the entire auditing and programming process can be interactive.

Finally, there is a great deal of interest both in academia and industry in systems that query "deep web" data sources, as well as systems that query ubiquitous sensors and other devices like network switches. Queries in these environments can be extremely long-running or even continuous, requiring online query processing techniques. These environments also present challenging unpredictabilities in performance. Because online query processing systems must handle unpredictable user control, the techniques developed for online query processing are coincidentally well-suited for these scenarios requiring adaptive query processing.

ACM SIGMOD 2001 May 21-24, Santa Barbara, California USA

2. TUTORIAL SKETCH

In this tutorial, we survey work on online query processing, including algorithms, statistical methods, system design issues and human-computer interactions. We review early work on database sampling and estimation, including a survey of current support in commercial products. We pose a set of performance metrics suited to matching the needs of users for online query processing systems. We look at online query processing algorithms including online reordering and ripple joins, and talk about how new estimation techniques and resulting measures of answer quality become an intimate part of the performance goals – and hence the design – of these algorithms. We also discuss architectural issues that arose in online query processing systems developed recently, including efforts in the core engines at both Informix and IBM. Finally, we place online query processing into context with other related work, including data reduction and data mining, precomputation schemes, as well as the notions of anytime algorithms in AI, and online algorithms in computational complexity.

3. REFERENCES AND NOTES

An annotated bibliography and an electronic version of the tutorial notes will be available via the website http://control.cs.berkeley.edu.