A Markovian Stream Warehousing System

Julie Letchner  Christopher Ré  Magdalena Balazinska  Matthai Philipose  Jennifer Wong

Lahar

A Markovian stream warehouse. It processes event and event-OLAP queries on Markovian streams.

**Motivating Scenario: Audio Search**

- **Q<sub>E</sub>:** Ed the economist wants to find all occurrences of the phrase "change in interest rates" in NPR newscasts from the last 6 months.
- **Q<sub>F</sub>:** FBI agent Fred wants to count the number of times the phrase "explosive device" appears in each phone conversation recorded (legally) in 2008.

**Problem:**
Ed and Fred are querying semantic content that is not explicitly expressed in their raw audio files. They can use speech recognition tools to produce transcripts, but because of noise and ambiguity, these transcripts are imprecise.

**Solution:**
Fred and Ed model these imprecise transcripts as Markovian streams and query them using Lahar. Lahar handles imprecision transparently by returning to Ed and Fred query results that are annotated with probabilities.

**What is a Markovian Stream?**
A Markovian stream is a compact representation of a probability distribution over sequences (here, sequences of words).

Markovian streams contain imprecision (drawn as boxes, e.g. 'rates' or 'freight') and temporal correlations (drawn as arrows).

**Event Query Processing**
Lahar's query processing algorithm:
- is implemented as a state machine (example below)
- follows standard possible worlds semantics
- runs in \(O(N^2D^2)\) time.
- single-pass

**Event-OLAP queries** aggregate the results of event queries (e.g. using COUNT).

**Processing Optimizations**
Lahar uses indexing and approximation techniques to accelerate query processing by orders of magnitude.

**Indexing**
Lahar leverages B+-tree indexes on search key <word, sequenceID>:

- \(t_{100}\)
- \(t_{101}\)
- \(t_{102}\)
- \(t_{103}\)

Output of event-OLAP query \(Q_F\):

**Approximation**
Lahar uses approximation to improve efficiency, creating an accuracy/efficiency trade-off. Approximations include:

- Independence: \(O(N^2D)\)
- Deterministic: \(O(N)\)

*compared to \(O(N^2D^2)\)*

**Demonstration Details**
Users interactively query audio streams and can verify results using playback.

**Data:**
- 20+ NPR news updates (~5 min each)
- Inferred using off-the-shelf speech processing

**Queries:**
- Pre-written queries for ease of use
- We encourage creation of original queries

**Can I use Lahar in my own work?**
Yes! Lahar will be released in September. Audio- and location-based Markovian stream data sets, along with source code, will be included in the release. Website address is below.

See Ré et. al. SIGMOD 2008 for details