Motivation

Web archives preserve the Web’s evolution—an important mirror of mankind’s recent history. Access to these collections is nowadays limited—a search functionality is often completely missing or ignores the temporal dimension.

In time-travel text search a keyword query is enriched by a temporal context (i.e., a time point or time interval). Only document versions that existed at any time in the temporal context are potential results of the time-travel query.

Data volumes in Web archives are easily at the order of Terabytes (more often Petabytes). This makes efficient time-travel text search a grand challenge!

Score Synopses

Relevance and scoring models for text search rely on numerous collection statistics that are time-varying in our setting. For time \( t \) we maintain, for instance:

- collection size \( N(t) \)
- average document length \( \text{avdl}(t) \)
- document frequency \( DF(v,t) \) of a term \( v \)
- PageRank score \( \text{pr}(d,t) \) of a document \( d \)

Typically, these collection statistics are sampled at a certain time granularity (e.g., monthly).

In [3] we proposed an efficient solution for the management of time-varying collection statistics. Individual statistics are viewed as a time series and a compact piecewise linear representation is obtained that retains a tunable approximation guarantee.

Score synopses provide the following benefits:

- **Space Economy:** Significant compression over the original representation is achieved.
- **Interpolation:** Collection statistics can be estimated for times in-between two observations.
- **Accuracy:** Reconstructed collection statistics are highly accurate.

Temporal Text Indexing

Standard text-indexing techniques do not provide efficient support for time-travel queries. Further, the high level of redundancy between document versions is not exploited. In [3] we proposed a temporal text-indexing technique having three key components:

- **Time-Travel Inverted File Index**
  We build on the inverted file index and extend the structure of its postings by a validity time-interval:

  \[ \langle d, \text{score}, [t_b, t_e] \rangle \]

- **Approximate Temporal Coalescing (ATC)**
  Subsequent document versions tend to have highly similar scores for the same term. ATC reduces the index size by coalescing such sequences, while retaining a tunable approximation guarantee.

- **Sublist Materialization (SM)**
  When processing a time-point query \( q \) many postings are read unnecessarily, since \( t \in [t_b, t_e] \). SM tackles this problem by systematically materializing smaller sublists. There are two variants of choosing sublists.

Performance Guarantee: The minimal amount of space is required, while retaining a tunable performance guarantee.

Space Bound: The best expected performance is achieved, while not requiring more than a tunable space bound.

Ongoing & Future Research

- Time-interval and fuzzy time-point queries
- Efficient temporal aggregations (as proposed in [2])
- Integration of positional information

References