SEDA: A SYSTEM FOR SEARCH, EXPLORATION, DISCOVERY AND ANALYSIS OF XML DATA

Andrey Balmin,† Latha Colby,† Emiran Curtmola,† Quanzhong Li,† Fatma Özcan,† Sharath Srinivas,‡ Zografoula Vagena,‡
†IBM Almaden Research Center †UC San Diego ‡UMD ‡Microsoft Research

Problem

• Extracting meaningful insights from heterogeneous XML data collections
  e.g. Find the average trade percentage amounts for import partners of United States

Challenges

• No fixed schema (driven by schema integration and evolution)
• Need to query both XML structure and text, and do not exactly know the structural constraints
• Need to compute analytics from semi-structured data (e.g., avg, min, max, count)
• Need 100% precision and recall for meaningful analytics, but keyword query results are imprecise

New Paradigm: Start with simple keyword search, employ user guidance to compute complex OLAP-style analytics

Keyword Search

• Maximum flexibility
• Queries are underspecified: hard to capture users' intentions
• Results are imprecise and ranked: inaccurate aggregates

SEDA System Overview

Keyword Search Query

Top-K Results

Keyword Search Query

Modified Query

Context Summaries (identify contexts)

Connection Summaries (identify relationships)

Top-K Results

Connection Summary

• Different relationships
  Let the user choose the meaningful connections between the query terms
  Infeasible to show all connections
  Compute and show only connections in top-K results, and exploit context filtering

Data Cube Computation

• Output is a table, one column for each query term,
  Consider query result as a de-normalized fact table
• Match each column in the result to a known dimension or measure
• Augment the query results with keys and values, as needed
• Compute the normalized dimension and fact tables
• Feed result into DB2 AlphaBlox® to compute aggregations and investigate

SEDA System Architecture

User Interface

Query Panel
Context Summary Panel
Connection Summary Panel
Result Panel
Data Cube Analysis Panel

Execution Engine

Top-k Join Processor
Context Summary Generator
Connection Summary Generator
Complete Result Set Generator
Data Cube Processor & Analyzer

Storage and Indexing

XML Data Storage
Full-text Index w/Keyword-path support
Reachability Bloom Filter Index
Edge Table
Data Guide Index

List of Dimensions

Name                          Context                     Key
Country                      /country                      e
Country                      /country/year               e
Country                      /country/trade_country       e
Partner Country              import_partners/ittrade_country   e
Partner Country              import_partners/pcttrade_country  e
Trade Amount                 /country/year, ...               e

List of Facts

Name                          Context                     Key
Country                      GDP                            e
Country                      GDP_ppp                        e
Country                      GDP_em                        e
Country                      GDP_ppp/em                    e
Country                      GDP_em/ppp                     e
Pattern Country              import_partners/ittrade_country   e
Pattern Country              import_partners/pcttrade_country  e
Trade Amount                 /country/year, ...               e

XML storage (DB2 pureXML®)

Indices
(full-text and structural)

Precise data, ready for summarization

Precise data, ready for summarization

OLAP-style data cubes

XML Data

List of Paths

XML storage (DB2 pureXML®)

Indices
(full-text and structural)

Precise data, ready for summarization

Precise data, ready for summarization

OLAP-style data cubes

XQuery, SQL, SQL/XML

Input query terms
• Context (tag name, path) and search term (Keywords)

Context Summary
(List of Paths)

Different paths may correspond to different real-world entities
  Let the user disambiguate paths

Connection Summary

• Different relationships
  Let the user choose the meaningful connections between the query terms
  Infeasible to show all connections
  Compute and show only connections in top-K results, and exploit context filtering

Data Cube Computation

• Output is a table, one column for each query term,
  Consider query result as a de-normalized fact table
• Match each column in the result to a known dimension or measure
• Augment the query results with keys and values, as needed
• Compute the normalized dimension and fact tables
• Feed result into DB2 AlphaBlox® to compute aggregations and investigate

Keywords

complex OLAP-style analytics
collections

SQL/XML

Search

e.g. Find the average amount for

import

USA

Return an initial set of possible answers
• Complex, hard to express
• Enable effective user interaction via user feedback loops
• Runtime discovery of XML contexts and connections between nodes
• Use relational data cube semantics to compute summarizations

Challenges

• No fixed schema (driven by schema integration and evolution)
• Queries are underspecified:
  Hard to capture users' intentions
• Results are imprecise and ranked: inaccurate aggregates
• Need 100% precision and recall for meaningful analytics, but
  keyword query results are imprecise