Efficient Querying of Distributed Linked Data

Martin Svoboda

svoboda@ksi.mff.cuni.cz

Faculty of Mathematics and Physics
Charles University in Prague
Czech Republic

25th March 2011

Outline

- Introduction
- Open problems
- Research intent
- Existing approaches
- Preliminary ideas
- Conclusion

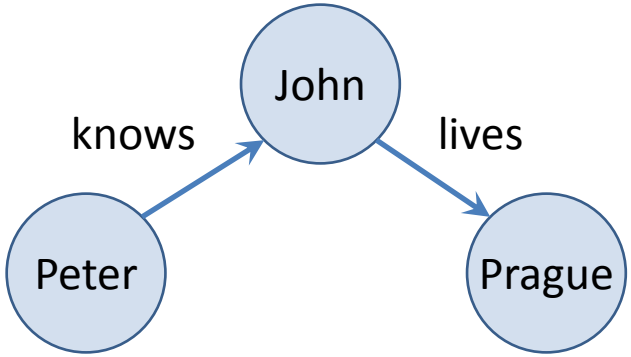
Introduction

- Motivation
  - Web of Documents
  - Web of Data
- Linked Data
  - Identifiers
  - Information
  - Links

Introduction

- RDF
  - Triples
    - Subject Predicate Object.
  - Graph
    - Vertices for subjects and objects
    - Edges for predicates

Peter knows John.
John lives Prague.



Open Problems

- Application architectures
  - Local/distributed approaches
  - Caching data copies
  - Accessing data online
- Links maintenance
  - Changes detection
  - Automatic correction

Open Problems

- User interaction
  - Query submitting
  - Data browsing
  - Links navigation
- Trust and quality
  - Data provenance
  - Social networks
  - Result ordering

Research Intent

- Querying framework
  - Index structures
  - Query processor
- Focused problems
  - Scalability
  - Distribution
  - Volatility

Research Intent

- Scalability
  - Large datasets
    - 200 datasets, 25 billions triples, 395 millions links
- Volatility
  - Values/links changes
- Distribution
  - Local/distributed approaches

Existing approaches

- Source selection
  - **Data Summaries for On-Demand Queries over Linked Data**
  - Andreas Harth et al.
- Query processor
  - **Matrix "Bit"loaded: A Scalable Lightweight Join Query Processor for RDF Data**
  - Medha Atre et al.

Source Selection

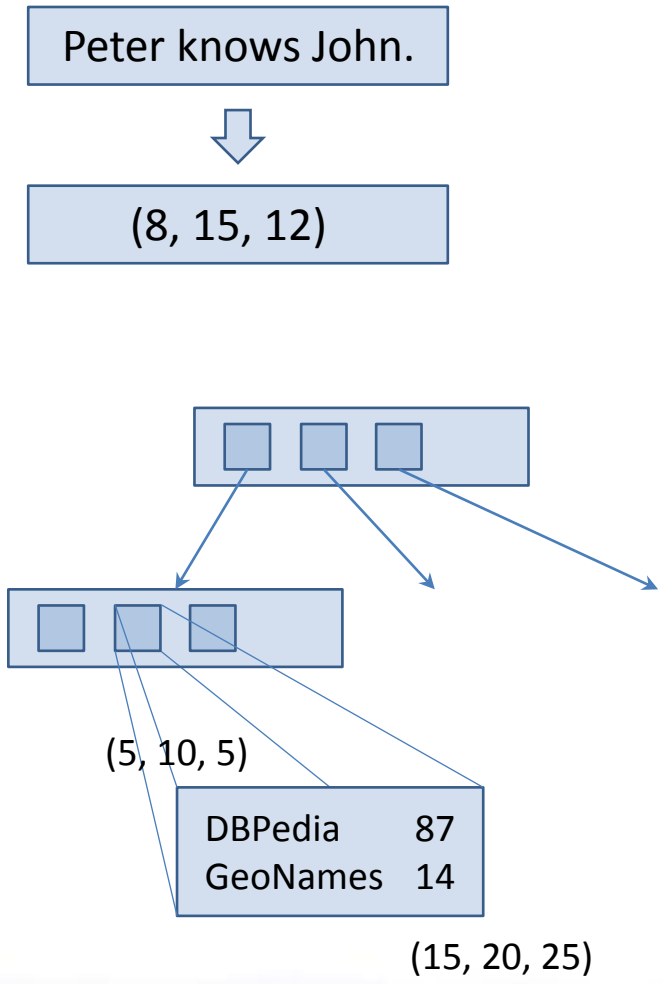
- Context
  - High number of small distributed datasets
  - Conjunctive SPARQL queries
- Contributions
  - Local index structure
  - Source selection algorithm

Peter knows John.
John lives Prague.

(Peter knows ?P)
(?P lives Prague)

Source Selection

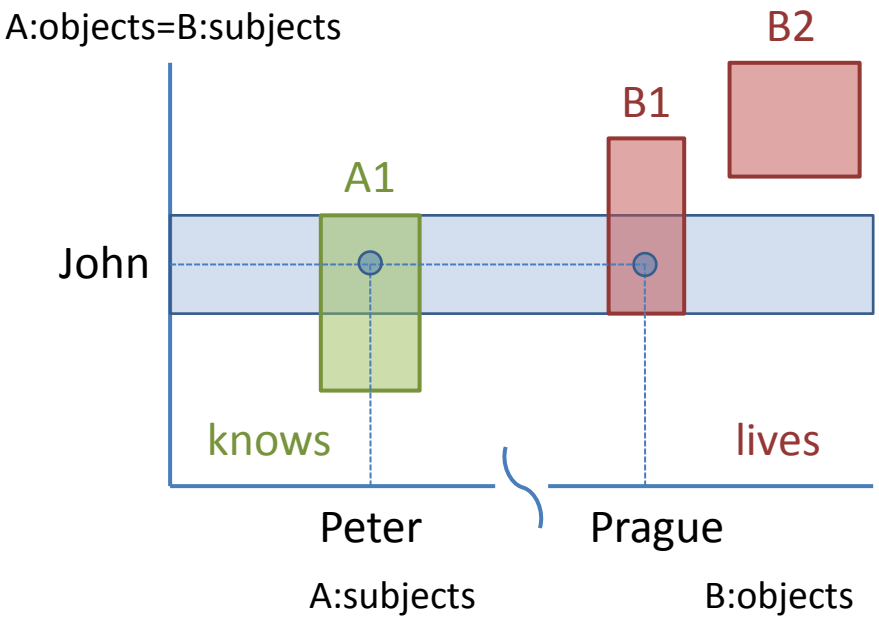
- Index structure
  - Data transformation
    - 3-dimenisonal space
    - Hash functions
  - Q-trees based on R-trees
    - Overlapping bounding boxes
    - Buckets with summaries



Source Selection

- Selection algorithm
  - Query transformation
  - Source selection
    - Individual patterns
      - Index traversal
      - Sets of buckets
    - Inductive joins
      - Required overlapping
  - Query processing

(Peter knows ?P)
(?P lives Prague)

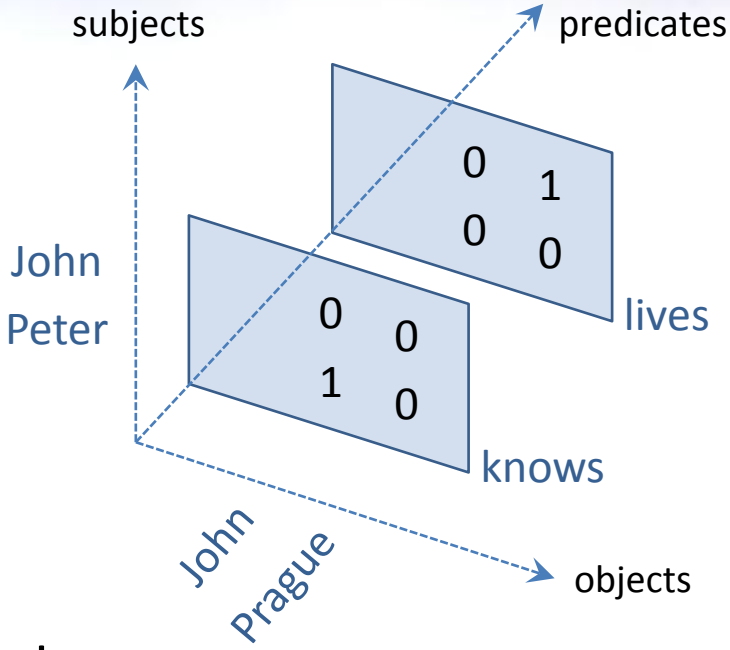


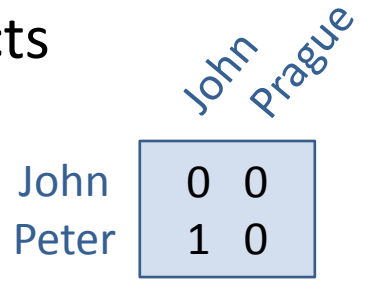
Query Processor

- Context
  - Local database
  - Conjunctive SPARQL queries
- Contributions
  - Index structure
  - Query evaluation algorithm

Query Processor

- Index structure
  - 3-dimensional matrix
    - S, P, O dimensions
    - Bit values
  - 2-dimensional slices
    - S-O, O-S slices for each predicate
    - P-O slices for subjects, P-S slices for objects
  - Physical file
    - Compressed bit runs



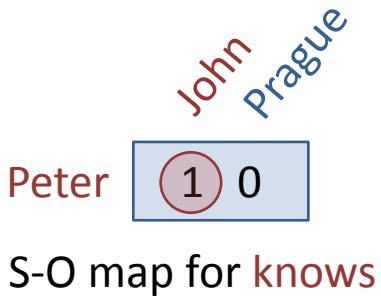


S-O slice for knows

Query Processor

- Evaluation algorithm
  - Initialization
    - Loading required parts of index
  - Pruning
    - Operations over compressed runs
  - Evaluation
    - Inductively joining graph patterns
    - Idea of nested loops algorithm

(Peter knows ?P)
(?P lives Prague)

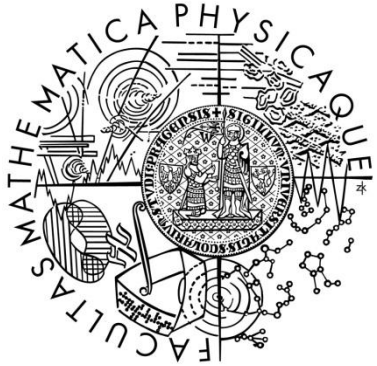


Preliminary ideas

- Ideas
  - Local dynamic index
  - Spatial techniques
  - Result ordering
- Plan
  - Real-world data analysis
  - Prototype implementation
  - Evaluation experiments

Conclusion

- Research area
  - Linked Data querying
    - Large/distributed/volatile datasets
- Planned contributions
  - Index structure
  - Query processor



Thank you for your attention…