Challenges in Big Data Management and Analytics Uncertainty, Structure, Intensionality

Pierre Senellart

3 February 2015, *SinFra 2015*

The Five Vs of Big Data

The Five Vs of Big Data

Volume: Data volumes beyond what is manageable by traditional data management systems (from TB to PB to EB)

The Five Vs of Big Data

Volume: Data volumes beyond what is manageable by traditional data management systems (from TB to PB to EB)

Variety: Very diverse forms of data (text, multimedia, graphs, structured data), very diverse organization of data

The Five Vs of Big Data

Volume: Data volumes beyond what is manageable by traditional data management systems (from TB to PB to EB)

Variety: Very diverse forms of data (text, multimedia, graphs, structured data), very diverse organization of data

Velocity: Data produced or changing at high speed (LHC: 100,000,000 collisions / second), more than able to store

The Five Vs of Big Data

Volume: Data volumes beyond what is manageable by traditional data management systems (from TB to PB to EB)

Variety: Very diverse forms of data (text, multimedia, graphs, structured data), very diverse organization of data

Velocity: Data produced or changing at high speed (LHC: 100,000,000 collisions / second), more than able to store

Veracity: Data quality very diverse; imprecise, imperfect, untrustworthy information

The Five Vs of Big Data

Volume: Data volumes beyond what is manageable by traditional data management systems (from TB to PB to EB)

Variety: Very diverse forms of data (text, multimedia, graphs, structured data), very diverse organization of data

Velocity: Data produced or changing at high speed (LHC: 100,000,000 collisions / second), more than able to store

Veracity: Data quality very diverse; imprecise, imperfect, untrustworthy information

Value: Making sense of potentially very valuable data, but with a value not immediately apparent

The Five Vs of Big Data

Volume: Data volumes beyond what is manageable by traditional data management systems (from TB to PB to EB)

Variety: Very diverse forms of data (text, multimedia, graphs, structured data), very diverse organization of data

Velocity: Data produced or changing at high speed (LHC: 100,000,000 collisions / second), more than able to store

Veracity: Data quality very diverse; imprecise, imperfect, untrustworthy information

Value: Making sense of potentially very valuable data, but with a value not immediately apparent

Special focus within IPAL: Web Data (Web pages, social networks, e-commerce data, Semantic Web, Linked Open Data, etc.)

Uncertain data is everywhere

Numerous sources of uncertain data:

- Measurement errors
- Data integration from contradicting sources
- Imprecise mappings between heterogeneous schemas
- Imprecise automatic processes (information extraction, natural language processing, etc.)
- Imperfect human judgment
- Lies, opinions, rumors

Structured data is everywhere

Data is structured, not flat:

- Variety of representation formats of data in the wild:
  - relational tables
  - trees, semi-structured documents
  - graphs, e.g., social networks or semantic graphs
  - data streams
  - complex views aggregating individual information
- Heterogeneous schemas
- Additional structural constraints: keys, inclusion dependencies

Intensional data is everywhere

Lots of data sources can be seen as intensional: accessing all the data in the source (in extension) is impossible or very costly, but it is possible to access the data through views, with some access constraints, associated with some access cost.

- Indexes over regular data sources
- Deep Web sources: Web forms, Web services
- The Web or social networks as partial graphs that can be expanded by crawling
- Outcome of complex automated processes: information extraction, natural language analysis, machine learning, ontology matching
- Crowd data: (very) partial views of the world
- Logical consequences of facts, costly to compute

Interactions between uncertainty, structure, intensionality

- If the data has complex structure, uncertain models should represent possible worlds over these structures (e.g., probability distributions over graph completions of a known subgraph in Web crawling).
- If the data is intensional, we can use uncertainty to represent prior distributions about what may happen if we access the data. Sometimes good enough to reach a decision without having to make the access!
- If the data is a RDF graph accessed by semantic Web services, each intensional data access will not give a single data point, but a complex subgraph.

Introducing UnSAID

- Uncertainty and Structure in the Access to Intensional Data
- Jointly deal with Uncertainty, Structure, and the fact that access to data is limited and has a cost, to solve a user's knowledge need
- Lazy evaluation whenever possible
- Evolving probabilistic, structured view of the current knowledge of the world
- Solve at each step the problem: What is the best access to do next given my current knowledge of the world and the knowledge need
- Knowledge acquisition plan (recursive, dynamic, adaptive) that minimizes access cost, and provides probabilistic guarantees

Introducing UnSAID

- Uncertainty and Structure in the Access to Intensional Data
- Jointly deal with Uncertainty, Structure, and the fact that access to data is limited and has a cost, to solve a user's knowledge need
- Lazy evaluation whenever possible
- Evolving probabilistic, structured view of the current knowledge of the world
- Solve at each step the problem: What is the best access to do next given my current knowledge of the world and the knowledge need
- Knowledge acquisition plan (recursive, dynamic, adaptive) that minimizes access cost, and provides probabilistic guarantees

Contributions to this project are welcome!