Provenance Incubator Group Teleconference
Provenance in (and among) databases

Note taker: Luc Moreau

26 Feb 2010

On Friday 26th, James Cheney organized a special session on "provenance in (and among) databases", with a few invited guests. Each guest presented their views on the subject for some ten minutes, followed by questions addressed to them.

1 Part 1: Susan Davidson

Professor Susan Davidson (University of Pennsylvania, http://www.cis.upenn.edu/~susan/) gave a brief introduction and then identified interesting research challenges in provenance for scientific workflows.

- Brief introduction to the state of the art:
  - A number of provenance challenges have taken place
  - Provenance challenges help investigate how systems can answer queries related to past execution
  - They have led to the specification of OPM (a de facto standard?), which is a good starting point in order to know how to represent provenance in scientific workflows
  - At the heart, OPM is a directed acyclic graph, which is leading to interesting questions.
- Research challenges
  - Distinguish workflow specification and execution; provenance is about execution.
  - Challenge to represent provenance successinctly. Some lines of work worth noting:
    * Collection oriented approaches e.g. Kepler
    * Recognising common substructures of provenance (Jagadish's team at U. of Michigan).
  - Figure out how to query provenance:
    * What is an appropriate query language for provenance?
    * Is it a graph query language?
    * There is another dimension: modules contain submodules/subworkflows. Hence, nesting has to be taken into account.
    * Query language for business process: Tel aviv group (Tova Milo)
    * Transitive closure queries can be expensive.

        - Taverna group and Pennsylvania labelling algorithms
    - How to deal with incompleteness of provenance?
    - Overload of provenance: provenance can be huge even for a small workflow
        - Idea of creating views, idea of composite module to hide part of provenance
    - Security and provenance: how do we know people have not fudged provenance?
        - Work by Winslett's team at U. of Illinois
        - Privacy concerns: if we can look at enough provenance, can we infer the workflow, which could be private.
            - So can we look at mechanisms to hide part of provenance to guarantee privacy?

- Questions
    - (Luc) Do notions of provenance in database and workflow communities converge?
    - (Yolanda) Yolanda found comment about privacy interesting. Very often privacy concerns are about keeping a user private (e.g. when a negative review about doctor is published: it may be important to preserve the author's privacy). Here, it is different, it's workflow privacy. It's about not giving a recipe away.
    - (Paulo) What can you automatically capture? How can scientists annotate information, e.g. parameters,etc?
        - Ontology tagging for workflow (Kepler, Vistrails, Taverna).

2 Part 2: Wang-Chiew TAN

Professor Wang-Chiew Tan (University of California, Santa Cruz, http://users.soe.ucsc.edu/~wctan/) talked about challenges related to provenance and annotations in databases.

- Challenges:
    - In data integration or data warehouse, it is essential to reason about provenance of data
    - Two approaches: lazy or eager.
        - lazy: don't keep anything
        - eager: they keep extra information as they transform data
    - Eager approach: what do you keep? what kind of provenance question can you answer? what should you keep minimally? how do you query that information?
    - Lazy approach: how do you query provenance?
    - How to apply provenance in various ways: for debugging rules, for transformation extraction rules in datawarehouse, for correcting (view maintenance problem)?
    - How to store/query/manipulate provenance effectively?
        - Requires identifying set of what to keep for what purpose.

- Questions
    - (Paul) What provenance is necessary to keep so that you could infer missing provenance (in an open world)?
      Why/Where/How notions of provenance: why is a piece of data in an output, where was it created, how was it created?

3 Part 3: Peter Buneman

Professor Peter Buneman (University of Edinburgh, http://homepages.inf.ed.ac.uk/opb/) presented his view about provenance.

- Peter's view on provenance:
  - Provenance is a huge problem;
  - Some 15 years ago, received attention from the database community;
  - We don't understand the problem, and even if it is a single problem;
  - In a database, problem can be illustrated as follows: a phone number shows up in a view, *why* is it there, *where* does it come from, *how* was it constructed?
  - Workflow provenance and database provenance seemed divergent.
    - Workflows have black boxes, grey bloxes, white boxes.
    - Database people also put black boxes in programming languages.
  - Ultimately, he believes this will converge in a single form of provenance.
  - He got involved in this because:
    - Curated databases are typically created by people copying data, possibly from other curated databases. So provenance problem is intrisic in curated databases.
    - But databases may have changed. Problem of maintaining/archiving older versions of the database.
    - Problem of data citation. We know how to cite papers, but how do we cite data in databases?
  - Semantic web community often makes a big deal about databases being "closed world". There are two issues. First, ever since database theory started, people have been looking at elements of the open world assumption. It is now very important in data exchange. (See Gottlob's Datalog+-). The second point is that it is not at all clear whether the distinction is important in thinking about provenance or temporal DBs.
  - In this context, it is also unclear RDF is the right starting point, given required extra column for named graph, extra column for time.
- Questions
  - (Luc) Can you apply where provenance to rdf-based query languages such as SparQL?
    - Yes, it's fairly trivial to apply this to SparQL.
    - Other problems with ontologies, how to deal with updates and time.
    - Peter would not start solving the problem from the notion of triple store, but from a more principled approach.
  - (Paul) where can we start work for tracking provenance on the web?
    - Peter "admires" xg-prov since it is very challenging!
    - Can we tackle the citation problem? it could be a starting point.
    - Is a standard committe the right way? Or is it better to have people "scratching their head" in their offices?

- (Yolanda) Some people say: why are you discussing provenance? where is the problem? Can't you just do it? What's Peter's answer to such questions?
  - It's challenging to identify what to actually record and do it effectively.
  - Very often, too much or too little:
    - If a workflow queries a database, we cannot capture the state of the database.
    - If some data is stored in a database, we are missing some context: what is the workflow we lost?
  - We need to understand how various models of provenance interact? The fundamental challenge is to obtain the right/decent model for provenance. We've made good progress but we are not there yet.

The teleconference ended at 5pm (GMT).

A Suggested Reading

1. Buneman, P. 2006. How to cite curated databases and how to make them citable. SSDBM 2006:195-203. http://portal.acm.org/citation.cfm?id=1155000
2. Buneman, P., Cheney, J., Tan, W., and Vansummeren, S. 2008. Curated databases. PODS 2008: 1-12. http://portal.acm.org/citation.cfm?id=1376918
3. Buneman, P. and Tan, W-C. Provenance in databases. SIGMOD 2007: 1171-1173. http://portal.acm.org/citation.cfm?id=1247646
4. Susan B. Davidson, Juliana Freire: Provenance and scientific workflows: challenges and opportunities. SIGMOD 2008:1345-1350 http://portal.acm.org/citation.cfm?id=1376616.1376772