Project ideas

CMPSCI 645

Spring 2010
Project requirements

**Form groups (~ Week 4)**
Find a project partner and begin to discuss project ideas.

**Project proposal (~ Week 6)**
Your proposal should explicitly state the following:
- Problem your project will address.
- Your project’s goal and motivation.
- Areas of related work. Your survey of the related work should place a particular focus on how existing work differs from your proposed work and why it is insufficient for solving the problem you propose. Through this survey, you should be able to convince the reader that you are addressing something fundamentally new, either a brand new problem or a novel approach to a known problem.
- The methodology and plan for your project. Be sure to structure your plan for the project as a set of incremental milestones and include a schedule for meeting them.
- The resources needed to carry out your project.

**Midterm status report (~ Week 10)**
Your status report should contain enough implementation, data, and analysis to show that your project is on the right track. You should revise your original proposal to accommodate the instructor’s comments, along with any surprising results or changes in direction, schedule, etc. You sometimes also need to have a refined version of the problem statement as well as a more developed related work section.

**Project presentation (Last 2 classes)**
A brief presentation should include the proposed problem, state-of-the-art solutions, your proposed solutions including the algorithms and implementation, and evaluation results. The presentation may include a system demo if appropriate.

**Project report (End of term)**
A final report extends your previous reports to present the research problem, summarize your contributions, survey related work, and include a detailed description of your algorithms, analysis, implementation, evaluation methodology and significant results, and finally present conclusions. For group work, the report should also include a paragraph explaining, for each group member, their contributions and duties in the project.
Large Scale Parallel Data Processing

- Investigate the use of compression in Hadoop.

- Investigate theoretical issues in MapReduce paradigm

- Investigate the implementation of recursive queries in MapReduce.

- Apply MapReduce/Hadoop to a large scale data analysis task of your choosing. Evaluate performance, expressiveness for task.

- Investigate databases reading from files without load phase.
Analyzing massive graphs

- It is now possible to collect and store massive graphs representing networks of contacts, interactions, or flows of information among entities.

- Network analysts seek to understand the structure and function of such graphs, but many graph analysis implementations assume graphs fit in memory. Relational systems not well-suited to graph analysis.

- Investigate alternative data analysis paradigms for large graphs:
  - Parallel processing for graph analysis (using Hadoop):
    
    **PEGASUS: A Peta-Scale Graph Mining System, ICDM 2009.**
  - For overview of network analysis tasks:
    
Databases and IR

• Enhanced provenance for information retrieval
  
  • How do we explain the presence (or absence) of search results?
  
  • How can a user compare the results of two search technologies?

• Access control for text/document databases
  
  • How can an access control policy be described and enforced over a complex text collection?

• See also: DBRank Workshop for recent work at the intersection of DB/IR
Deterministic v. Probabilistic Databases

- Conventional databases are deterministic:
  - An item is either in the database, or it is not.
  - A tuple is either in the query answer, or it is not.

- Probabilistic databases:
  - “An item belongs to the database” is a probabilistic event
  - “A tuple is an answer to the query” is a probabilistic event
### Text extraction

**address string**  “52-A Goregaon West Mumbai 400 076”

<table>
<thead>
<tr>
<th>House</th>
<th>Area</th>
<th>City</th>
<th>Pincode</th>
<th>Prob</th>
</tr>
</thead>
<tbody>
<tr>
<td>52</td>
<td>Goregaon West</td>
<td>Mumbai</td>
<td>400 062</td>
<td>0.1</td>
</tr>
<tr>
<td>52-A</td>
<td>Goregaon</td>
<td>West Mumbai</td>
<td>400 062</td>
<td>0.2</td>
</tr>
<tr>
<td>52-A</td>
<td>Goregaon West</td>
<td>Mumbai</td>
<td>400 062</td>
<td>0.5</td>
</tr>
<tr>
<td>52</td>
<td>Goregaon</td>
<td>West Mumbai</td>
<td>400 062</td>
<td>0.2</td>
</tr>
</tbody>
</table>

Four segmentations of the address string along with probabilities
Probabilistic database ideas

• Focus on a specific domain. Develop representation mechanisms and query answering techniques.

• For an existing probabilistic database model, develop an interface for presenting answers to users.

• Develop probabilistic representations for trust and authenticity in collaborative authoring.

Reliability in Integrated / Collaborative data sources
Classical databases

- Strict control over data creation and update
- Known users
  - registered, authenticated
  - access control mechanisms in DBMS or interfaces limit their actions
- Data sources accurate
- Queries and query answers are precise.
These assumptions often fail

- Uncertainty about information is often unavoidable in these applications
  - Data integration frameworks
  - Web extraction systems
  - Collaborative authoring platforms
  - Scientific databases
Gerome Miklau

Publications: (14) (total 75 citations)
Sorted by date | citations

2005
- Nilesh Dalvi, Gerome Miklau, Dan Suciu. Asymptotic Conditional Probabilities for Conjunctive Queries. ICDT, 2005 (1 citation)
- Gerome Miklau, Dan Suciu. Containment and equivalence for a fragment of XPath. J. ACM vol 51, pages 2, 2004 (3 citations)
- Gerome Miklau, Dan Suciu. Modeling Integrity in Data Exchange. Secure Data Management, 2004 (0 citations)
- Gerome Miklau, Dan Suciu. A Formal Analysis of Information Disclosure in Data Exchange. SIGMOD Conference, 2004 (2 citations)

2004
- Gerome Miklau, Dan Suciu. Controlling Access to Published Data Using Cryptography. VLDB, 2003 (4 citations)
- Todd J. Green, Gerome Miklau, Makoto Onizuka, Dan Suciu. Processing XML Streams with Deterministic Automata. ICDT, 2003 (21 citations)
- Igor Tatarinov, Zachary G. Ives, Jayant Madhavan, Alon Y. Halevy, Dan Suciu, Nilesh Dalvi, Xinyi Dong, Yana Kadiyska, Gerome Miklau, Peter Mork. The Piazza peer data management project. SIGMOD Record vol 32, pages 47, 2003 (13 citations)

2002
- Gerome Miklau, Dan Suciu. Containment and Equivalence for an XPath Fragment. PODS, 2002 (27 citations)
- Gerome Miklau, Dan Suciu. Cryptographically Enforced Conditional Access for XML. WebDB, 2002 (0 citations)

????
- Gerome Miklau, Dan Suciu. Containment and Equivalence of Tree Patterns. UW Technical Report TR 02-02-03, Computer Science Department University of Washington (3 citations)
Collaborative authoring

Authors

Collaborative source

End users

Author 1

Author 2

Author 3

Alice

Bob

Charlie

updates

updates

updates

Thursday, February 4, 2010
Problem: many sources of inaccuracy

• Data sources inaccurate, incomplete, or inconsistent

• Integration/extraction technologies imperfect

• Autonomous authors

  • Mistaken or malicious behavior

• Weak assurances of identity: anonymous or pseudonymous

• Numerous authors: contradictory actions

How can end-users evaluate the quality of information presented to them?
A user’s reliability questions

- Is this data value accurate?
- Are data values missing?
- Which authors/sources contributed to this data value?
- If I have beliefs in the trustworthiness of sources or authors, how does that impact my assessment?
  - What influence does author/source X have over this data value?
  - Improve implementation of belief database: “Believe it or not: Adding belief annotations to databases”.
  - Investigate the authenticity of information in Wikipedia.
Protecting sensitive information
Protecting online content

• Valuable content is presented freely on the Web

• Legitimate users usually retrieve a small set of pages

• What stops one from stealing entire websites and replicating them elsewhere?

• Choose a domain. Investigate techniques for resisting content theft while not interfering with legitimate users.

  • Audit requests and identify malicious use?

  • Make crawling hard?
Privacy

- Investigate the accuracy of a data mining or learning algorithm under differential privacy.

Study Google cluster data

Google recently released cluster data and posed a number of problems:

- How do we characterize workloads?
- How do we generate predictive models of workloads?
- How do we assign tasks to machines optimally?

http://googleresearch.blogspot.com/2010/01/google-cluster-data.html
Network traces

- What is the state of the art in network trace analysis?

- Can database systems improve performance?

- Focus on
  - querying traces
  - transforming traces

- Should this be treated as a stream processing problem?

- Can traces be usefully indexed?
Not enough project ideas?

- Talk to Prof. Diao or Prof. McGregor about possible projects.