Crossing the Structure Chasm


Presentation by:
Abram Hindle

Department of Computer Science
University of Waterloo
ahindle@cs.uwaterloo.ca

November 21, 2005

This Presentation

• What am I going to cover?
  – Authors
  – Introduction
  – Motivation
  – Definitions
  – U-World / S-World
  – Property of Web Data
  – Revere, Mangrove, Piazza, DesignAdvisor/MatchAdvisor
  – Future Work
  – Summary
  – Discussion

Authors

• From University of Washington
• Alon Y. Halevy, Professor, Computer Science and Engineering
• Primary Researcher - Interests data access in heterogeneous environments, Schema Matching, Machine Learning
• Oren Etzioni, Professor, Director of the Turing Center (Semantic Web) - Interests Semantic Web, etc.
• Anh Hai Doan, Assistant Professor at Siebel Center for Computer Science, University of Illinois - Schema Matching, Data Int.
• Zachary G. Ives, Assistant Professor at Computer & Information Science Department, University of Pennsylvania - Databases and data sharing
• Jayant Madhavan, PhD in Computer Science and Engineering at the University of Washington - Schema Mapping
Introduction

Chasm Exists Between Unstructured Data and Structured Data

Unstructured Data:
- Web pages
- Documents
- Human Created Information that lacks a Schema

Structured Data:
- Database relations
- Schemas
- Data with a Schema

Motivation

Bridging the chasm will allow for:
- Easier Annotation of unstructured data
- Use keyword search in a structured domain
- Ease of content Creation
- Accurate Searches and Aggregation
- Was the Chasm created by the tools?
- Can it be fixed or bridged with tools?
**U-World**
- Natural Language
- Unstructured Data
- Easy to Author
- Easy to Search
- Inaccurate to query
- Change Resistant
- Predominant form of information on-line
- No schema knowledge

**S-World**
- Fits a Schema
- Structured Data
- Hard to Author
- Hard to Search
- Accurate Queries
- Weak to Change
- Usually found in Deep Web.
- Schema knowledge required

**Property of Web Data**
- Web data is in HTML
- Hyper-Links
- Markup
- Could be further marked up
- Sometimes difficult to parse
- Flexible
- Generally Unstructured, has layout structure
Revere

- Annotate HTML with Schemas
- Share data via Schema Transitive Mapping
- P2P System
- Promote pro-annotation feedback cycles.
- Enable aggregation of annotated data

Mangrove

- HTML Annotation
- Schema creation and Matching
- Choose appropriate schema
- Annotations as RDF
- Positive Feedback Loop
  - Encourage users to annotate their HTML with a schema
Figure 1: Search Query: "assistant professor" <facultyMember> <portrait> ?

Figure 1: Search Query: "assistant professor" <facultyMember> <portrait> ?
Piazza

- P2P Data Management System
- Mediates the schema between each peer.
- No Global Schema
- Transitive Schema Mapping
- XML Based
- Data Sharing, answering, storage
- Use XQuery to aggregate various data sources.

Figure 2: PDMS for our university example. The arrows correspond to schema mappings between peers. No central mediated schema is necessary. As long as the mapping graph is connected, any peer can access data at any other peer by following schema mapping "links".

[HED+03]

Fig 4. Matching a query template into a target graph of a schema mapping. The matching tree pattern are shown in bold. The scheme mapping corresponding to the ontology graph is shown on the right.

[HIMT03]
Statistics

- Use TD/IDF in S-World
- Corpus of structures: OO, XML, DTDs, Ontologies
- Known schema mappings
- Actual data: tables, XML docs, ground facts of knowledge-base
- Queries over schemas and ontologies
- Basic Stats: term usage, co-occurring schema elements, similar names
- Composite Stats: composites
- Used by DesignAdvisor and MatchingAdvisor

DesignAdvisor/MatchingAdvisor

- Use the stats generated from the various corpuses.
- DesignAdvisor
  - Finds Similar Schemas
  - Allows making new schemas based on old ones
  - \( \text{sim}(S', (S, D)) = \alpha \cdot \text{fit}(S', S, D) + \beta \cdot \text{preference}(S') \)
  - \( \alpha \) and \( \beta \) are weights
  - \( \text{fit} \) measure the fit for \( S \) and \( S' \), ratio between mappings and total # of elements
  - \( \text{preference} \) measures conformity or common usage of \( S' \)

MatchingAdvisor

- MatchingAdvisor
  - Uses Schema Matching and Mapping Techniques like LSD or GLUE
  - Machine Learning based Mapping
  - Alternatively can use DesignAdvisor for matching and ranking
Future Work

- Future papers explore Piazza and Mangrove
- Deeper Discussion of Piazza and Mangrove
- Research proves groundwork for quite a bit of future work.
- Directions include transitive schema maps
- Intelligent Data Placement
- Distributed Querying

Summary

- U-World and S-World Semantics
- Describe how data is used in each world
- Suggest a system to overcome it.
- Use Sociological feedback reinforcement argument
- P2P Data Management with Aggregation

Discussion

- Was the chasm bridged?
- Does the chasm exist in the data, the schemas, the tools or not at all?
- Are we evolving these features or are we choosing better features: RSS, Blogs, Tagging, Web2.0
- Tool Adoption - Would anyone actually use their tools?
- What are alternative ways of supporting annotation or creating structured data?
- What are the effects of partially annotated data on this system?
- Is it safe to assume that we can transitively map all the schemas?
References

