Crossing the Structure Chasm

Alon Halevy, Oren Etzioni, AnHai Doan,

Zachary Ives, Jayant Madhavan, Luke McDowell,

Igor Tatarinov

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.
- AnHai 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

- Luke McDowell, Assistant Professor at Dept. of Computer Science, United States Naval Academy - Semantic Web, Email
- Igor Tatarinov, Phd in Computer Science and Engineering at the University of Washington - Data-mining, Schema Mediation
- Context
  - In First Biennial Conference on Innovative Data Systems Research
  - Preliminary Work
  - 31 Citations of Paper (over 7 of which are self references)
  - Followed up by papers on the subsystems of Revere

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

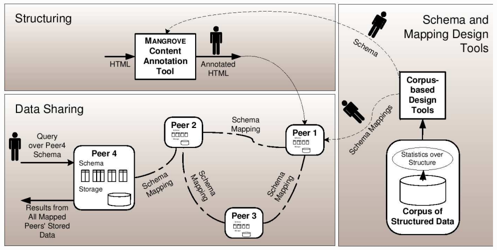
- Fits a Schema
- Structured Data
- Hard to Author
- Hard to Search
- Accurate Queries
- Weak to Change
- Usually found in Deep Web.
- Schema knowledge required

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


Structuring
HTML
MANGROVE Content Annotation Tool
Annotated HTML
Schema and Mapping Design Tools
Schema
Corpus-based Design Tools
Schema Mappings
Statistics over Structure
Corpus of Structured Data
Data Sharing
Query over Peer4 Schema
Peer 4
Schema
Storage
Results from All Mapped Peers' Stored Data
Peer 2
Schema Mapping
Peer 1
Schema Mapping
Schema Mapping
Schema Mapping
Peer 3

[HED+03]

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

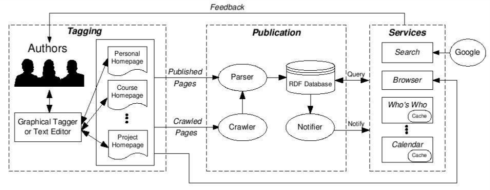
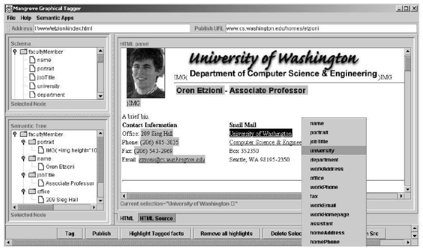


Figure 1: The MANGROVE architecture and sample services.

[MEG+]



Address l:\www\etzioni\index.html

Publish URL www.cs.washington.edu/homes/etzioni

```
facultyMember
  name
  portrait
  jobTitle
  university
  department
```

```
facultyMember
  portrait
    IMG(<img height="10
  name
    Oren Etzioni
  jobTitle
    Associate Professor
  office
    209 Sieg Hall
```

University of Washington

IMG(Department of Computer Science & Engineering)IMG

Oren Etzioni - Associate Professor

)IMG

A brief bio.

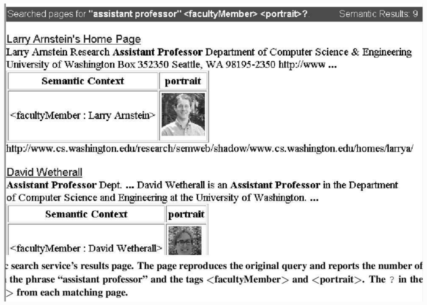
Contact Information
Office: 209 Sieg Hall
Phone: (206) 685-3035
Fax: (206) 543-2969
Email: etzioni@cs.washington.edu

Snail Mail
University of Washington
Computer Science & Enginee
Box 352350
Seattle, WA 98195-2350

Current selection="University of Washington □"

Figure 3: The MANGROVE graphical tagger. The pop-up box presents the set of tags that are valid for tagging the highlighted text. Items in gray have been tagged already, and their semantic interpretation is shown in the "Semantic Tree" pane on the lower left. The user can navigate the schema in the upper left pane.

[MEG+]



Searched pages for **"assistant professor" <facultyMember> <portrait>?** Semantic Results: 9

Larry Arnstein's Home Page
Larry Arnstein Research **Assistant Professor** Department of Computer Science & Engineering University of Washington Box 352350 Seattle, WA 98195-2350 http://www ...

Semantic Context	portrait
<facultyMember : Larry Arnstein>	

http://www.cs.washington.edu/research/semweb/shadow/www.cs.washington.edu/homes/larrya/

David Wetherall
Assistant Professor Dept. ... David Wetherall is an **Assistant Professor** in the Department of Computer Science and Engineering at the University of Washington. ...

Semantic Context	portrait
<facultyMember : David Wetherall>	

: search service's results page. The page reproduces the original query and reports the number of the phrase "assistant professor" and the tags <facultyMember> and <portrait>. The ? in the > from each matching page.

Figure 1: Search Query: "assistant professor" <facultyMember> <portrait> ?
[MEG+]

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.

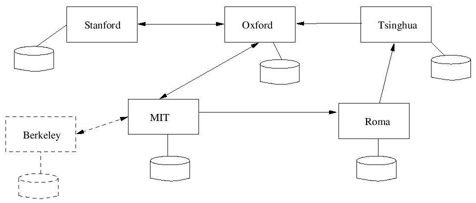
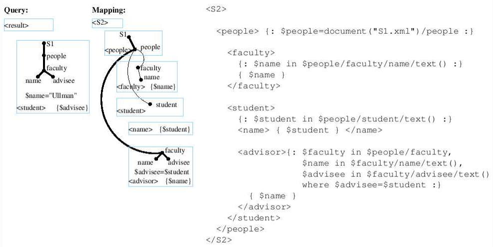


Figure 2: PDMS for our university example. The arrows correspond to schema mappings between peers. No central mediated schema is necessary. As long as the mapping graph is connected, any peer can access data at any other peer by following schema mapping "links".

[HED+03]



```
<S2>

  <people> {: $people=document("S1.xml")/people :}

    <faculty>
      {: $name in $people/faculty/name/text() :}
      { $name }
    </faculty>

    <student>
      {: $student in $people/student/text() :}
      <name> { $student } </name>

      <advisor>{: $faculty in $people/faculty,
                  $name in $faculty/name/text(),
                  $advisee in $faculty/advisee/text()
                  where $advisee=$student :}
        { $name }
      </advisor>
    </student>
  </people>
</S2>
```

Fig. 4 Matching a query tree pattern into a tree pattern of a schema mapping. The matching tree patterns are shown in bold. The schema mapping corresponding to the middle 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
  - $sim(S', (S, D)) = \alpha\dot{f}it(S', S, D) + \beta\dot{p}reference(S')$
  - α and β are weights
  - fit measure the fit for S and S', ratio between mappings and total # of elements
  - 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 ground work 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 feature or are we choosing better feature: 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

[HED$^+$03] A. Halevy, O. Etzioni, A. Doan, Z. Ives, J. Madhavan, L. McDowell, and I. Tatarinov. Crossing the structure chasm, 2003.

[HIMT03] A. Halevy, Z. Ives, P. Mork, and I. Tatarinov. Piazza: Data management infrastructure for semantic web applications, 2003.

[MEG$^+$] Luke McDowell, Oren Etzioni, Steven D. Gribble, Alon Halevy, Henry Levy, William Pentney, Deepak Verma, and Stani Vlasseva. Evolving the semantic web with mangrove.