Structure and Dynamics of Emergent Semantics Systems

Karl Aberer

EPFL School of Computer and Communication Science

NCCR MICS, National Centre of Competence in Research on Mobile Information and Communication Systems

karl.aberer@epfl.ch

www.mics.org

lsirwww.epfl.ch
Overview

1. Emergent Semantics
2. Mapping Inference in Semantic Overlay Networks
3. Structure of Semantic Overlay Networks
4. Peer Data Management Systems Implementation
5. Outlook: Sensor Internet
Semantics

- Long-standing debate: “What is semantics?”
- Standard response: “Mapping of a syntactic structure into a semantic domain”

Syntactic structure: database, knowledge base
Semantic domain: real-world
Semantic Web

- Real-world is a somewhat ill-defined and hard to compute concept
- Proposal: Substitute real-world by shared formal conceptualization [Gruber 93]

Syntactic structure: database, knowledge base
Semantic domain: ontology
The Issue with Ontologies

• What is the semantics of ontologies? After all, they are syntactic structures!
• Heterogeneous and evolving models and ontologies
The Correspondence Continuum

- Observation: Meaning is rarely a simple mapping from a syntactic structure to a semantic domain
- Continuum of (semantic) correspondences from symbol to (symbol to)* object [Smith 87]
- The meaning of a symbol is given by the composition of the semantic mappings that relate it to its root
- Instead of focusing on ever-richer modelling languages for concepts, focus on mapping languages and mapping discovery tools [Mylopoulos 06]
Semantic Grounding

- Still: what should “relate ot the root mean”?
- The meaning of symbols can be explained by its semantic correspondences to other symbols alone [“Understanding understanding” Rapaport 93]

- Type 1 semantics: understanding *in terms of something else*
  - Problem: how to ground semantics?
- Type 2 semantics: understanding something *in terms of itself*
  - “syntactic semantics”: grounding through recursive understanding
The National Centres of Competence in Research are a research instrument of the Swiss National Science Foundation.

**Emergent Semantics**

- Semantic correspondences form rich self-referential structures: recursive understanding.
- Recursive understanding may converge against stable structures (fixpoints): *emergent semantics* [Aberer 04]

![Diagram of emergent semantics]

- owl:tiger
- oil:cat
- wn:animal
- rdf:tiger
Peer-to-Peer Systems

- Resource Sharing (e.g. images)
  - No centralized infrastructure
  - Global scale information systems
  - Application-specific overlay networks
Efficiently Searching Resources (Data)

- Find images taken last week in Trondheim!
Resource Sharing

- What is shared?

knowledge

content

bandwidth

processing

storage
Beyond Keyword Search

- Support of structured data at peers: schemas
- Structured querying in peer-to-peer system
Peer Data Management Systems

Q1 =
<GUID>$p/GUID</GUID>
FOR $p IN /Photoshop_Image
WHERE $p/Creator LIKE "%Robi%"

Photoshop
(own schema)

<Photoshop_Image>
   <GUID>178A8CD8865</GUID>
   <Creator>Robinson</Creator>
   <Subject>
      <Bag>
         <Item>
            Tunbridge Wells
         </Item>
         <Item>Royal Council</Item>
      </Bag>
   </Subject>
</Photoshop_Image>

T12 =
<Photoshop_Image>
   <GUID>$fs/GUID</GUID>
   <Creator>
      $fs/Author/DisplayName
   </Creator>
</Photoshop_Image>
FOR $fs IN /WinFSImage

WinFS
(known schema)

<WinFSImage>
   <GUID>178A8CD8866</GUID>
   <Author>
      <DisplayName>
         Henry Peach Robinson
      </DisplayName>
      <Role>Photographer</Role>
   </Author>
   <Keyword>
      Tunbridge
   </Keyword>
   <Keyword>Council</Keyword>
   ...
</WinFSImage>

Q2 =
<GUID>$p/GUID</GUID>
FOR $p IN T12
WHERE $p/Creator LIKE "%Robi%"

⇒ Extending data integration techniques to decentralized settings
Semantic Heterogeneity in PDMS

- Pairwise mappings
  - Local mappings overcome global heterogeneity
  - Iterative query reformulation
PDMS vs. Classical Data Integration

- Traditional database techniques (e.g., LAV/GAV) rely on centralized schemas to integrate data sources

\[ m(\text{myDate}) = \text{Date} \quad m(\text{yourDate}) = \text{Date} \]

- Not applicable to large-scale, decentralized contexts
  - Scale: 100s vs. \(10^3-10^6\)
  - Churn: no fixed topology
  - Autonomy: no transactions, no integrity constraints, no global schema
Emergent Semantics in PDMS

- P2P data management systems form (complex) mapping networks between models:
  **Semantic Overlay Networks (SON)**
  - Mappings manually or automatically generated
  - Mappings establish semantic correspondences
  - Mutually negotiated and verified (pragmatic dimension)
- Practical systems with the potential to exhibit emergent semantics properties
- Technical challenges?
Overview

1. Emergent Semantics
2. Mapping Inference in Semantic Overlay Networks
3. Structure of Semantic Overlay Networks
4. Peer Data Management Systems Implementation
5. Outlook: Sensor Internet
The National Centres of Competence in Research are a research instrument of the Swiss National Science Foundation.

**Answering Queries in PDMS**

- **Semantic Query routing**
  - To whom shall I forward a query posed against my local schema?
- **Some (most) mappings will be (partially) faulty**
  - Different views on conceptualizations
  - Low expressive power of mapping languages
  - Automatic schema matching techniques
- **Alternatives**
  - Local query resolution only: Low recall
  - Flooding the whole network (PDMS so far): Low precision
Analyzing PDMS Mapping Networks

- **Standard deductive inference is not sufficient**
  - Uncertainty on mappings and conceptualizations
  - Precision/Recall tradeoff

- **Analyze Mapping Network**
  - Transitive closures of mapping operations
  - Cycles and parallel paths
  - Check for consistency
  - **Abductive reasoning**: find best possible explanation in case of inconsistency
Example

$q: \text{art/Creator?}$

$q \, \text{VS} \, m_3(m_4(m_0(q)))$

art/Creator? VS art/creatDate?
The National Centres of Competence in Research are a research instrument of the Swiss National Science Foundation

**Probabilistic Message Passing (Semantic Gossiping)**

- Deriving **quality measures** for the mappings using feedback
  - Reduces uncertainty
  - Used to route query / optimize mappings

![Diagram of Probabilistic Message Passing](image)
Using feedback

- The result of applying a composite mapping to a query should be identical to the original query for a cycle
  - Allows to estimate the probability of correctness of mapping

\[ P(f_0^+ | m_0, \ldots, m_{n-1}) = \begin{cases} 
1 & \text{if all mappings correct} \\
0 & \text{if one mapping incorrect} \\
\Delta & \text{if two or more mappings incorrect}
\end{cases} \]
Computing a Marginal for One Cycle

unknown observed

• $P(m_0, m_3, m_4, f_0) = P(m_0) \cdot P(m_3) \cdot P(m_4) \cdot P(f_0 | m_0, m_3, m_4)$

• Determine $P(m_i | f_0)$ given $P(f_0 | m_0, m_3, m_4)$?

• $P(m_0 | f_0) = \sum_{m_3, m_4} P(m_0, m_3, m_4, f_0) \cdot P(f_0)^{-1}$

• But: feedbacks on different cycles are correlated
  – One wrong mapping will affect several cycles/paths
  – Need to express a global probabilistic model for the mapping graph
A Brief Intro to Factor-Graphs

- \( g(x_1, x_2, x_3, x_4) = f_A(x_1, x_2)f_B(x_2, x_3, x_4) \)

\[
g_2(x_2) = \left( \sum_{x_1} f_A(x_1, x_2) \right) \left( \sum_{x_3} \sum_{x_4} f_B(x_2, x_3, x_4) \right)
\]
Deriving PDMS Factor-Graphs

Innate probabilities of mappings being correct

\[ P(f^+_\text{circ} | m_0, \ldots, m_{n-1}) \]
PDMS Factor-Graphs

- Cyclic graph
  - Junction Tree?
    - Centralization
    - Computational + communicational overhead
  - Iterative Sum-Product
    - Correct only for tree structured networks
    - Approximate result
- How to perform iterative sum-product by message passing on the mapping graph?
  - Message passing in factor graph does not correspond to connectivity of mapping graph
  - We want to rely on decentralized computations only
Embedded Message Passing

- Decentralized computations
  - Computationally inexpensive
  - Sums and Products
- Message-Passing Schedules
  - Lazy (piggybacking on query forwarding)
    - No message overhead
  - Periodic

local message from factor $f_{a_j}$ to mapping variable $m_i$:
$$
\mu_{f_{a_j} \rightarrow m_i}(m_i) = \sum_{m_j} \left( f_{a_j}(X) \prod_{p_k \in n(f_{a_j})} \mu_{p_k \rightarrow f_{a_j}}(p_k) \prod_{m_l \in n(f_{a_j}) \setminus \{m_i\}} \mu_{m_l \rightarrow f_{a_j}}(m_l) \right)
$$

local message from mapping $m_i$ to factor $f_{a_j} \in n(m_i)$:
$$
\mu_{m_i \rightarrow f_{a_j}}(m_i) = \prod_{f_{a_k} \in n(m_i) \setminus \{f_{a_j}\}} \mu_{f_{a_k} \rightarrow m_i}(m_i)
$$

remote message for factor $f_{a_k}$ from peer $p_0$ to peer $p_j \in n(f_{a_k})$:
$$
\mu_{p_0 \rightarrow f_{a_k}}(m_i) = \prod_{f_{a_l} \in n(m_i) \setminus \{f_{a_k}\}} \mu_{f_{a_l} \rightarrow m_i}(m_i)
$$

Posterior correctness of local mapping $m_i$:
$$
P(m_i | \mathcal{F}) = \alpha \left( \prod_{f_{a_k} \in n(m_i)} \mu_{f_{a_k} \rightarrow m_i}(m_i) \right)$$
Evaluation: Convergence

(undirected example graph, prior 0.7 delta 0.1)
Evaluation: Fault-tolerance (faulty links)

(undirected example graph, prior 0.8 delta 0.1)
Evaluation: Performance Detecting Errors

(randomly generated networks of 50 schemas and 200 mappings, TTL = 5)
Conclusion Probabilistic Message Passing

• A technique to implement emergent semantic processes
  – Decentralized decision making
  – Converges to an agreement on conceptualizations
  – Scalable and robust method to infer correct mappings in a semantic overlay network

• Questions
  – Do such mapping networks exist?
  – What is their structure?
Overview

1. Emergent Semantics
2. Mapping Inference in Semantic Overlay Networks
3. Structure of Semantic Overlay Networks
4. Peer Data Management Systems Implementation
5. Outlook: Sensor Internet
Semantic Overlay Networks

- Networks of schema mappings
  - Directed, weighted, redundant
- Semantic Interoperability
  Two peers are said to be *semantically interoperable* if they can forward queries to each other in the mapping graph, potentially through series of translation links
- Question
  - Which are necessary conditions that a semantic overlay network becomes semantically interoperable in the large-scale?
- Idea: use percolation theory to detect the emergence of a strongly connected component in $S$
  - Adaptation of a recent graph-theoretic framework (Newman, Strogatz, Watts 2001)
The Model

- Large-scale semantic overlay networks as random graphs with arbitrary degree distribution
- Specificities of the model
  - Strong clustering (clustering coefficient cc)
  - Bidirectionality (bidirectionality coefficient bc)
- Based on *generatingfunctionology* ($p_k$ probability of degree k)

\[
G_0(x) = \sum_{k=0}^{\infty} p_k x^k
\]

- Necessary condition for semantic interoperability in the large
  - Appearance of a giant strongly-connected component: $c_i > 0$

\[
c_i = \sum_k k(k-2-cc)p_k
\]
The National Centres of Competence in Research are a research instrument of the Swiss National Science Foundation.
Evaluation: The Sequence Retrieval System (SRS)

- Bioinformatic libraries: EMBL, SwissProt, Prosite, etc.
  - Commercial information indexing and retrieval system
  - Links from one database to others by mapping identifiers
  - More than 380 databanks and 500 (undirected) links
  - Custom crawler
Results

- Connectivity indicator $ci = 25.4$
  - Giant connected component: 187 nodes
- Size of the giant component
  - 0.47 (derived)
  - 0.48 (observed)
- Powerlaw Topology
- Small-World Graph
  - Clustering coefficient = 0.32
  - Diameter = 9

\[ y(x) = \alpha x^{-\gamma} \text{ with } \alpha = 0.21 \text{ and } \gamma = 1.51 \]
Graphs with Same Power-law Degree Distribution

- Varying number of edges
Analyzing Weighted Networks

- Do we have a sufficient number of good mappings?
- Using quality measures from the mappings derived from message passing
  - Uniformly distributed weights between 0 and 1
  - Attribute / schema level
- Semantic query forwarding
  - Per-hop forwarding behaviors
  - Only forward if $w_i \geq \tau$
    - $\tau = 0$: flooding
    - $\tau = 1$: exact answers
Overview

1. Emergent Semantics
2. Mapping Inference in Semantic Overlay Networks
3. Structure of Semantic Overlay Networks
4. Peer Data Management Systems Implementation
5. Outlook: Sensor Internet
GridVine: Annotating Shared Resources

- **End-users** create annotations / "categories" / "translation links"
  - Constraining the annotation mechanism: we do not expect them to write ontologies, views...
GridVine: Annotating Shared Resources

- Principle of data independence
  - Scalable physical layer: structured overlay network (P2P network)
  - Semantic logical layer: Semantic Gossiping
Mapping annotations onto P-Grid

P-Grid: Structured overlay network
- Supports key-based search
- Decentralized, scalable, self-organizing access structure

User-defined annotations (RDF triples)

User-defined category translations (OWL)

User-defined categories (RDFS)

⇒ RDQL queries
Traversals of the Semantic Overlay Network

- **GridVine**: structured P2P network = Distributed index
  - Query forwarding independent of structure of semantic overlay!

- **Different query forwarding paradigms**
  - Iterative forwarding
  - Recursive forwarding

![Graph](image.png)
Overview

1. Emergent Semantics
2. Mapping Inference in Semantic Overlay Networks
3. Structure of Semantic Overlay Networks
4. Peer Data Management Systems Implementation
5. Outlook: Sensor Internet
Outlook

Information Sharing in the Sensor Internet

(Source: activecampus/UCSD)
Current Situation with Information Sharing

Different WSNs

Web publishing (repetitive)
DB app, java app, Web interface, ...

Discovery and correlation (difficult)
Syntactic and semantic heterogeneity
Challenges

- Provide generic and simple-to-use tools for *publishing* data collected from sensors over the Web
  - Data stream management

- Provide tools for *discovering* published sensor data
  - Semi-structured metadata

- Provide tools for *correlating* data from autonomous and heterogeneous sensor data sources
  - Emergent semantics
Publishing: Global Sensor Network

- A simple-to-use system to publish and correlate sensor data streams
  - Virtual sensors are sensors, sensor networks or derived data streams
  - GSN nodes managing virtual sensors
- Architecture
  - Virtual Sensors published in a P2P network using metadata annotations
  - GSN nodes connected in a peer-to-peer streaming network in the Web
  - Data processing specified in a temporal SQL extension

Example

4 Motes (Light and Temperature)[last 10 seconds], 1 RFID reader[last 1 value], 2 Wireless Cam[last 10 seconds]

SELECT camera WHERE
  (AVG(Temperature) > 30 OR
  RFIDReader.value NOT IN
  (SELECT TAG_ID
   FROM personel
   WHERE personel.lab = LSIR ) )
Development Status

Available through sourceforge: http://sourceforge.net/projects/globalsn
Discovery: PicShark

- Assume sensor data is being massively published
  - Already the case for images (photo sharing, Flickr!)
- Discovery depends on the availability of annotations
  - Content-based search capabilities are limited (no text!)
- Manual annotations are hard to obtain
  - Metadata scarcity
- Social annotation (folksonomies)
Exploiting Social Context

- Assume standardized annotation scheme
  - Attributes $A_1, ..., A_n$
- Information-theoretic measure of metadata scarcity of image $I$

\[
H(I) = -\sum p(A_i) \log(p(A_i))
\]

where $p(A_i) = \begin{cases} 
0 & \text{if attribute is present} \\
\frac{1}{n} & \text{otherwise (n : #attributes)}
\end{cases}$

- Reducing metadata scarcity by metadata propagation to similar images
  - Similarity derived from metadata, features, user relevance feedback
PicShark Prototype

- Extract existing metadata in different formats
- Propagate metadata
- Extract features from image content and text annotations
- Store metadata in standard formats in a peer-to-peer network

The National Centres of Competence in Research are a research instrument of the Swiss National Science Foundation
Summary Information Sharing

- Publishing and sharing of sensor data in a peer-to-peer architecture: Global Sensor Network

- Shared annotations of image/sensor data in a social network: PicShark

- Distributed reasoning on the correctness of mappings among heterogeneous annotation schemes: emergent semantics
The National Centres of Competence in Research are a research instrument of the Swiss National Science Foundation.
Acknowledgements

- Joint research with the following members and visitors of my group
  - Philippe Cudre Mauroux, Adriana Budura, Ali Salehi, Manfred Hauswirth, Andras Feher, Tim van Pelt, Julien Gaugaz

- Funding provided by
  - Swiss National Foundation (SNF) through NCCR MICS
  - European Union through the Evergrow project (FET)
For more information

- www.mics.ch
- lsirwww.epfl.ch
- www.p-grid.org
- sourceforge.net/projects/globalsn

Thanks for your attention!