Building RDF- and Schema-Based Peer-to-Peer Systems

Wolfgang Nejdl

L3S / University of Hannover
Germany
Overview

Relevant L3S Project Background
- Motivation
- Project Background - PADLR, Edutella, et al

Schema-Based Peer-to-Peer Networks
- Characteristics and Building Blocks
- Resource Description Framework (RDF) and RDF Schema
- Edutella Query Service / RDF Query Exchange Language RDF-QEL
- Semantic Web Inferencing
- Subscriptions
- Efficient Routing / HyperCuP & Super-Peers
- Integration of new Peers / Clustering
- Distributed Query Processing in P2P Networks
- Mediation

Summary and Conclusions
Motivation

Distributed Peer-to-Peer Infrastructures for the Semantic Web

Semantic Web Metadata Standards for describing (learning) resources and users

Personalized Environments in the Adaptive Web

How can we use distributed (learning) resources in a personalized way?
PADLR and Edutella

Personalized Access to Distributed Learning Repositories
(www.learninglab.de/english/projects/padlr.html)

Most important (CS) modules
- Peer-to-Peer Infrastructure (incl. Clients and Providers)
- Courseware Watchdog and Metadata Extraction
- Personalization and Personalized Queries based on Metadata
- Web-(Learning-) Services

PADLR Participants
(Stanford, Hannover, Karlsruhe, Stockholm, Uppsala)

+ Edutella Participants
(Vienna, Berlin, Darmstadt, etc.)
Edutella: Goal and Approach

Specify and implement a RDF-based meta-data infrastructure for P2P networks

Developed as part of the open source peer-to-peer project JXTA edutella.jxta.org

> 60 contributors from various institutions
EU/FP6 NoE KnowledgeWeb

Scalability

Heterogeneity

Languages

Dynamics

Semantic Web Services
EU/FP6 NoE PROLEARN

Working towards
- innovative elearning resources
- interoperable elearning resources and systems
- sustainable elearning infrastructures and processes for SMEs
EU/FP6 NoE REWERSE

Reasoning on the Web with Rules and Semantics

Develop reasoning languages for advanced Web systems

Test these languages on adaptive Web systems and Web-based decision support systems

Bring these languages to the level of pre-standards

Selected applications for proof-of-concept purposes

Personalized Web systems

Web-based decision support

Towards a Bioinformatics Semantic Web
Overview

Relevant L3S Project Background
- Motivation
- Project Background - PADLR, Edutella, et al

Schema-Based Peer-to-Peer Networks
- Characteristics and Building Blocks
- Resource Description Framework (RDF) and RDF Schema
- Edutella Query Service / RDF Query Exchange Language RDF-QEL
- Semantic Web Inferencing
- Subscriptions
- Efficient Routing / HyperCuP & Super-Peers
- Integration of new Peers / Clustering
- Distributed Query Processing in P2P Networks
- Mediation

Summary and Conclusions
### Schema-Based Peer-to-Peer Networks

- **User-definable schemas**
- **Structured schemas**
- **Query language**
- **Decentralized control**
- **Node autonomy**
- **Transient peers**
- **Self organization**

#### Database Systems
- **ANY RDBMS**
- **CONCEPTBASE**
- **ONTOBROKER**
- **AMOSII**
- **OBJECTGLOBE**
- **TSIMMIS**
- **TUKWILA**

#### Schema-based P2P Systems
- **CHATTY WEB**
- **EDUTELLA**
- **PIAZZA**

<table>
<thead>
<tr>
<th></th>
<th>Database Systems</th>
<th>Schema-based P2P Systems</th>
</tr>
</thead>
<tbody>
<tr>
<td>fixed</td>
<td>ANY RDBMS</td>
<td>CHATTY WEB</td>
</tr>
<tr>
<td>schema/keywords</td>
<td>CONCEPTBASE</td>
<td>EDUTELLA</td>
</tr>
<tr>
<td>key</td>
<td>ONTOBROKER</td>
<td>PIAZZA</td>
</tr>
<tr>
<td>local</td>
<td>AMOSII</td>
<td>fixed schema/keywords</td>
</tr>
<tr>
<td>distributed</td>
<td>OBJECTGLOBE</td>
<td>key</td>
</tr>
<tr>
<td>peer-to-peer</td>
<td>TSIMMIS</td>
<td>fixed schema/keywords</td>
</tr>
<tr>
<td></td>
<td>TUKWILA</td>
<td>key</td>
</tr>
<tr>
<td></td>
<td></td>
<td>fixed schema/keywords</td>
</tr>
</tbody>
</table>

(System list not complete)
Building Blocks

Flexible Schema Language
  - to describe complex and heterogeneous resources in the P2P network

Expressive Query Language
  - to retrieve data from heterogeneous data stores

Efficient Network Topology
  - to allow appropriate routing algorithms

Mediation Facilities
  - to integrate and combine (possibly heterogeneous) information
RDF / RDF Schema for Describing Distributed Resources

Basic Formalisms for the Semantic Web

- URIs to identify resources
- Combine resources and annotate resources with attributes, using `<Subject, Property, Value>` Tuples
- Graph as basic model, easy to translate to logic facts
- RDFS allows us to define the RDF vocabulary used (classes and attributes), and thus to represent simple semantic models
- Possible extensions towards more expressive semantic descriptions, e.g. description logic (DAML+OIL / OWL)

Using RDF / RDFS in the P2P context

- Distributed annotations for distributed resources
- Flexible schema definitions, which can be uniquely identified and combined, as well as extended by additional properties
Characterization of Peers using RDFS

Schema level
- Supporting specific schemas: dc, lom, dcq

Property level
- Supporting specific properties: dc:subject, lom:type, dc:format

Property value range
- Supported ranges for specific properties, e.g. ccs:dbms for dc:subject

Property values
- Specific attribute values, e.g. „exercise“ for lom:type, „en“ for dc:language
RDF-QEL: RDF Query (Exchange) Language

Datalog-based Query Exchange Language (RDF-QEL)

- RDF QEL1: conjunctive query up to
- RDF QEL5: RDF QEL4 (SQL3) + general recursion

see Nejdl et al: “EDUTELLA: A P2P Networking Infrastructure Based on RDF“, WWW 2002

- Datalog is used as the internal data model (ECDM: Edutella Common Data Model) and provided as a set of Java classes
- RDF is used to represent the queries transmitted between the peers
- Wrappers for other RDF query languages (RQL, TRIPLE, etc.) and XML query languages (like Xpath)
From Querying to Reasoning

World Wide Web Data as Distributed (Web) content
+ Semantic Web Metadata
Distributed and interoperable (RDF) metadata descriptions about:
- Content
- Relationships between the content
- Learner
+ Semantic Web Inferencing
(Logic) Programs and Rules to:
- Adapt the content and relationships (links)
- Infer new metadata
= Declarative and Composable Web Services
siehe auch REWERSE NoE

P2P

Content
Content Metadata
Relationships
Learner Model
Logic Programs
P2P and Semantic Web Inferencing: Edutella as basic infrastructure for ELENA (EU/FP5)
Another Possibility: Don‘t query, subscribe

Subscriptions are a good idea, too (get the NYTimes each morning, get new teaching material on P2P topologies ...)

Example: Selective Information Dissemination in P2P-DIET

Instead of Queries and Answers we need

- Profile forwarding
- Notification forwarding / Filtering
- Advertisement forwarding
- Dynamicity of P2P network → storing notifications / rendezvous

See e.g. Koubarakis et al: Selective Information Dissemination in P2P Networks: Problems and Solutions, SIGMOD Record, Special P2P Issue, September 2003
P2P and Efficient Routing

How do peer-to-peer networks scale?

Requirements:

- Symmetric topology (every node is a root)
- Low network diameter (small worlds property, should be $O(\log n)$)
- Limited node degrees (number of peer-connections from a node, should be $O(\log n)$)
- Load balancing of traffic
- Efficient broadcast (receive broadcast messages only once)
- Adaptable to dynamic number of peers
HyperCuP Peer-to-Peer Topology

Details: see e.g. Schlosser, Sintek, Decker, Nejdl: „HyperCuP - Shaping Up Peer-to-Peer Networks“, 2nd Intl. WS on Agents and P2P Computing, 2002
Hypercube Topology

Broadcast Algorithm
- Annotate messages with the “dimension” of the peer-to-peer connection, and only forward it along “higher” dimensions

Properties
- Network diameter, characteristic path length and number of nodes are $O(\log_b N)$
- Fault tolerant, vertex-symmetric
Super-Peer Networks

Observation: Peers vary significantly in availability, bandwidth, processing power, etc.
Create network backbone from highly available and powerful peers to distribute load better.
See also Yang, Garcia-Molina: Improving Search in P2P Systems, Intl. Conf. on Distributed Computing Systems, Vienna, 2002, or file sharing networks like KaZaa
Super-Peers and Routing Indices

Nejdl et al. Super-Peer-Based Routing and Clustering Strategies for RDF-Based Peer-To-Peer Networks. WWW 2003
Extension to Distributed Query Processing

Interleave P2P techniques and query processing
- Push abstract query plans through the super peer network
- Super peers pick and expand those parts of the query plan that can be executed locally
- On the fly distribution and expansion of query plans

See Brunkhorst, Dhraief, Kemper, Nejdl, Wiesner: Distributed Queries and Query Optimization in Schema-Based P2P-Systems, VLDB-P2P-Workshop
Clustering for Better Routing

Have to use Clustering to make routing indices efficient

- Query-Based Clustering: cluster on query and data characteristics, using frequency counting algorithms to identify the most relevant item sets to be included in the indices and to be used for clustering

- Rule-based clustering: cluster based on user-specified rules (cmp. DirectConnect and E-Donkey file sharing networks), which explicitly state the clustering criteria (see Löser, Nejdl, Wolpers, Siberski: Information Integration in Schema-Based P2P Networks, CAISE 2003)
Mediation: Some P2P-Specific Issues

Which basic assumptions should we take?

- Peer databases: relational databases or deductive databases based on Datalog (and definitively with minimal model property)
  - Motivation: Moving from key-word based P2P systems to schema-based systems is a good idea for more general P2P information systems, but these schema-based systems should not be too complicated
- No global schema, but mapping rules between two peers: range-restricted rules with conjunctive queries in body and head
  - Motivation: These or simpler mapping rules are probably not too difficult to create (a P2P system might need many of them), and they take care of the dynamicity of the P2P environment
- „global program“ can again be seen as a Datalog program
  - see Franconi et al, VLDB-P2P-Workshop
Mediation: Some P2P-Specific issues

Further observations

- Acyclical mapping rules might actually be sufficient (see also Piazza: Halevy et al, WWW 2003 (Peer data management systems: Infrastructure for the Semantic Web), ICDE 2003)

- Cycles in the mapping rules might not be meant as recursions, but could be used for checking the quality / completeness of mapping rules (see also Aberer et al: The Chatty Web, WWW 2003)

- Mapping / matching vocabularies might be sufficient, too (see He, Chang: Statistical Schema Matching across Web Query Interfaces, SIGMOD 2003)
Overview

Relevant L3S Project Background
- Motivation
- Project Background - PADLR, Edutella, et al

Schema-Based Peer-to-Peer Networks
- Characteristics and Building Blocks
- Resource Description Framework (RDF) and RDF Schema
- Edutella Query Service / RDF Query Exchange Language RDF-QEL
- Semantic Web Inferencing
- Subscriptions
- Efficient Routing / HyperCuP & Super-Peers
- Integration of new Peers / Clustering
- Distributed Query Processing in P2P Networks
- Mediation

Summary and Conclusions
Summary and Conclusions

Schema-based P2P networks and P2P-based data management infrastructures build upon traditional P2P networks and distributed / heterogeneous database research, while posing new challenges as well as additional functionalities.

Building blocks are flexible / extendable schema languages (such as RDFS), expressive query and reasoning languages, efficient network topologies as well as routing and clustering algorithms, and finally advanced, but not too complicated, data integration and mediation functionalities.

See also SIGMOD Record September 2003, Special P2P Issue: Nejdl, Siberski, Sintek: „Design Issues and Challenges for RDF- and Schema-Based Peer-to-Peer Systems“