Ontology-based Clustering in a Peer Data Management System

Ph.D. Thesis

Carlos Eduardo Santos Pires (cesp@cin.ufpe.br)
Advisor: Ana Carolina Salgado (acs@cin.ufpe.br)

April 27, 2009 - Recife, PE, Brazil
Introduction

Peer Data Management Systems (PDMS)
Problem

- Arbitrary approach for connecting peers is inefficient
- Peers sharing
  - Different content (exported schemas)
    - Neighbors in the overlay network
  - Similar content
    - Positioned far from each other or even isolated in the overlay network
Motivation

- **Semantic Communities**
  - Put together peers with common interests about a specific topic
  - Formulated queries are transmitted among the peers of the community
  - Peers are organized according to a P2P network topology
  - Exported schemas represented by ontologies
  - Creation and maintenance is a challenging issue
Goal

• Main Contribution
  – A process for clustering peers into the semantic communities that compose a PMDS

• To achieve this objective, we propose…
  – Ontology-based PDMS architecture
  – Ontology matching process
    • Global Similarity measure
  – Automatic process for summarizing ontologies
  – Peer Clustering Process
Background

- Ontology Matching and Merging
Background

• Clustering
  – Automatic process of partitioning a finite set of objects in a set of meaningful clusters
  – Exclusive and unsupervised classification

• Clustering Issues (Peer Clustering)
  – Object set availability
  – Sensitivity to input order

• Cluster Validity
  – External and Internal
SPEED: Semantic Peer Data Management System

DHT Network

SP1

SP2

SP3

IPi1

DPi11

DPi12

IPi2

DPi21

DPi22

IPij

DPij1

DPij2

Unstructured Super-Peer Network

Semantic Peer
Integration Peer
Data Peer
Semantic Community
Semantic Cluster

Ontology-based Clustering in a Peer Data Management System
Ontologies in SPEED

- **Semantic Peer**
  - Community Ontology
  - Summarized Cluster Ontology

- **Integration Peer**
  - Cluster Ontology
  - Local Ontology

- **Data Peer**
  - Local Ontology

- **Cluster**
  - Community
  - Community Ontology

- **Local Ontology**
  - Cluster Ontology
  - Local Ontology

- **Data Peer**
  - Local Ontology

- **Ontology-based Clustering in a Peer Data Management System**
Other Definitions

• Requesting Peer
  – Peer wishing to join the system
  – Connected as a data peer or integration peer

• Semantic Neighbor (Cluster)
  – Belong to the same community
  – Share semantically similar content

• Semantic Neighborhood
  – Set of semantic neighbors of a cluster
Data Peer

Ontology-based Clustering in a Peer Data Management System
Integration Peer
Semantic Peer

Ontology-based Clustering in a Peer Data Management System
Architectural Considerations

• Why a DHT network?
  – Efficient searches and sensibility to changes in the structure
  – Semantic Peers
    • High reliability, network bandwidth, and availability

• Why a peer takes part in only one cluster?
  – Avoid duplication of query results
Architectural Considerations

• Why a super-peer network?
  – Provides an environment that is better suited to the establishment of schema mappings
  – Facilitate query routing
  – Avoid multiple successive reformulations
  – Exploit the physical heterogeneity of peers

• Why a semantic index?
  – Avoid starting the search for a semantically similar cluster in an ad-hoc manner
# SPEED vs. Related PDMS

<table>
<thead>
<tr>
<th>PDMS</th>
<th>Network Topology</th>
<th>Network Population</th>
<th>Domains</th>
<th>Neighborhood Search</th>
<th>Semantic Similarity</th>
</tr>
</thead>
<tbody>
<tr>
<td>OntSum</td>
<td>Unstructured</td>
<td>Not empty</td>
<td>Predefined</td>
<td>Flooding; short and long distance links</td>
<td>Ontology Matching</td>
</tr>
<tr>
<td>Sunrise</td>
<td>Unstructured</td>
<td>Empty</td>
<td>Non existing</td>
<td>Centralized Access Point Structure (APS); SCI</td>
<td>Distance between concepts</td>
</tr>
<tr>
<td>Helios</td>
<td>Unstructured</td>
<td>Not empty</td>
<td>Non existing</td>
<td>Flooding</td>
<td>Ontology Matching</td>
</tr>
<tr>
<td>SPEED</td>
<td>Mixed</td>
<td>Empty</td>
<td>Predefined</td>
<td>Semantic index and flooding</td>
<td>Ontology Matching</td>
</tr>
</tbody>
</table>
Matching Process

1. **Linguistic-Structural Matching** (any matcher)

2. **Semantic Matching**

3. **Similarity Combination**

4. **Correspondence Ranking**

5. **Correspondence Selection**

Phase 1: 1:n or n:m

Phase 2: 1:1

Ontology \( O_i \)

Domain Ontology Weights

Semantics Rules Application

Ontology \( O_j \)

Weights

\( A_{LS} \)

\( A_{SE} \)

\( A_{CO} \)

\( A_{ij} \)
Example (Semiport and UnivBench)
Global Similarity Measure

Alignment $A_1$

1. Person, Person, isEquivalentTo 1.0
2. FullProfessor, FullProfessor isEquivalentTo 1.0
3. UndergraduateStudent, Course, isPartOf 0.3
4. Student, Person, isSubConceptOf 0.8
5. Professor, Faculty, isSubConceptOf 0.8

Alignment $A_2$

1. Person, Person, isEquivalentTo 1.0
2. FullProfessor, FullProfessor isEquivalentTo 1.0
3. Course, UndergraduateStudent, isWholeOf 0.3
4. Worker, Person, isSubConceptOf 0.8
5. GraduateStudent, UndergraduateStudent, isDisjointWith 0.0
6. Faculty, Professor, isSuperConceptOf 0.8
7. MasterStudent, Student, isSubConceptOf 0.8

$\text{Weighted Average}(O_i, O_j) = \frac{(1.0 + 1.0 + 0.3 + 0.8 + 0.8) + (1.0 + 1.0 + 0.3 + 0.8 + 0.0 + 0.8 + 0.8)}{16 + 17} = 0.66$
Implementation Issues

1. Choose ontology 1:
   - C\semport.owl

2. Choose ontology 2:
   - C\ UniBench.owl

3. Choose domain ontology:
   - C\ EM001\Education.owl

Semantic matching:
- Execute semantic matching

Similarity:
- Calculate SSM
- Choose the function to use: Average/DICE

Save to...
- DB/OWL

Ontology-based Clustering in a Peer Data Management System
Experiments

Recall

Precision

\[ R(A, R) = \frac{|R \cap A|}{|R|} \]

\[ P(A, R) = \frac{|R \cap A|}{|A|} \]
OWLSum: an Ontology Summarization Process

- Main use in Peer Clustering
  - Resume cluster ontologies (semantic index)
- A summary does not represent a cluster ontology in its entirety
  - Improve ontology matching
Relevance Measures

- **Centrality**: relationships (number and type) of a concept with other concepts in an ontology $O$

  $$\text{centrality}(c_n) = \frac{nr \times \left( \frac{n_s \times w_s}{\max_s} + \frac{n_{ud} \times w_{ud}}{\max_{ud}} \right)}{|C| - 1}$$

- **Frequency**: occurrences of a concept in local ontologies $O_1, \ldots, O_n$ that compose $O$

  $$\text{frequency}(c_n) = \frac{|\text{correspondences}(c_n)|}{|O_1, \ldots, O_n|}$$
Summarization Process

1. Ontology O
   - Calculate Concept Relevance
   - Determine Relevant Concepts

2. Group Adjacent Relevant Concepts
   - [Verify Groups of Concepts]
   - 0, 1 (without all relevant concepts), or n Groups of Concepts
   - Identify Paths Between Relevant Concepts
   - Analyze Identified Paths
   - Determine Ontology Summary

1 Group of Concepts with all relevant concepts

Ontology Summary OS
Example

Ontology Summary

<table>
<thead>
<tr>
<th>Concept</th>
<th>Relevance</th>
<th>Centrality</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>ServerSoftware</td>
<td>0.231</td>
<td>0.077</td>
<td>0.077</td>
</tr>
<tr>
<td>Cable</td>
<td>0.192</td>
<td>0.115</td>
<td>0.115</td>
</tr>
<tr>
<td>NetworkNode</td>
<td>0.192</td>
<td>0.115</td>
<td>0.115</td>
</tr>
<tr>
<td>SwitchEquipment</td>
<td>0.192</td>
<td>0.115</td>
<td>0.115</td>
</tr>
<tr>
<td>Computer</td>
<td>0.192</td>
<td>0.115</td>
<td>0.115</td>
</tr>
<tr>
<td>SecurityEquipment</td>
<td>0.192</td>
<td>0.115</td>
<td>0.115</td>
</tr>
<tr>
<td>RoutingComputer</td>
<td>0.192</td>
<td>0.115</td>
<td>0.115</td>
</tr>
<tr>
<td>NodePair</td>
<td>0.077</td>
<td>0.077</td>
<td>0.077</td>
</tr>
</tbody>
</table>

Recall = 100%  Precision = 86%  F-measure = 92.5%  Size = 7  Relevance Average = 0.181

Recall = 100%  Precision = 86%  F-measure = 92.5%  Size = 7  Relevance Average = 0.187

Ontology-based Clustering in a Peer Data Management System
## Experiments

<table>
<thead>
<tr>
<th>conference.owl</th>
<th>4-Concept</th>
<th>8-Concept</th>
</tr>
</thead>
<tbody>
<tr>
<td>Expert 1 vs. OWLSum</td>
<td>75%</td>
<td>88%</td>
</tr>
<tr>
<td>Expert 2 vs. OWLSum</td>
<td>50%</td>
<td>75%</td>
</tr>
<tr>
<td>Expert 3 vs. OWLSum</td>
<td>75%</td>
<td>75%</td>
</tr>
<tr>
<td>User agreement vs. OWLSum</td>
<td>75%</td>
<td>75%</td>
</tr>
<tr>
<td>User agreement vs. OntoSum</td>
<td>50%</td>
<td>50%</td>
</tr>
</tbody>
</table>
Peer Clustering in SPEED

DHT Network

Provide Local Ontology
[knowledge domain]

Search for Semantic Community

/ community not found
no connection

/ community found

SPEED Clustering Process
Unstructured Super-Peer Network

Search for Initial Cluster in Semantic Index
[initial cluster]

Search for most Similar Cluster

/ cluster found

Insert into most Similar Cluster

/ cluster not found

Create New Cluster

connection

Ontology-based Clustering in a Peer Data Management System
Search for a Semantic Community
Clustering Algorithm

- Inspired in the Leader algorithm [Hartigan, 1975]
- Main steps
  - Step 1. Search for Initial Cluster in Semantic Index
  - Step 2. Search for Most Similar Cluster
  - Step 3. Connection of a Requesting Peer
Step 1. Search for Initial Cluster in Semantic Index

- Requesting Peer $R_{P_n}$ sends its local ontology $L_{O_n}$ to semantic peer
- Search in the Semantic Index
  - $\text{SemMatch} (L_{O_n}, O_{S_{ij}})$
  - For each index entry a global similarity measure is produced
  - Initial cluster $\rightarrow$ highest global measure
Step 2. Search for Most Similar Cluster

Initial Cluster

Direct Neighbors

ConnectTTL = 3

IP_{i1}: 0.2
IP_{i2}: 0.6
IP_{i3}: 0.7
IP_{i4}: 0.4
3) Connection of a Requesting Peer

- **Case 1: MAX(global measure) ≥ cluster threshold**
  - RPₙ joins chosen cluster (Data Peer)
  - Merge (CLOᵢⱼ, LOₙ)
  - Semantic index is updated

- **Case 2: otherwise**
  - RPₙ creates new cluster (Integration Peer)
  - CLOᵢⱼ = LOₙ
  - A new entry is added to the semantic index
  - Semantic Neighborhood: Neighbor Threshold

![Graph showing neighbor and cluster thresholds]

IPᵢ₁, IPᵢ₄, IPᵢ₂, IPᵢ₃
Maintenance Considerations

- Evolution of cluster ontologies
  - Peer connection and disconnection
- Peer Disconnection
  - Removal of elements and semantic mappings
- Update of Cluster Neighborhood
- Recalculation of Global Similarity Measure
  - Similarity is calculated when a requesting peer joins a cluster
Implementation Issues

- **SPEED Simulator**
  - Implementation: Java
  - Integrated with *SemMatch* and *OWLSum*
  - Ontology Library (*Education*)

- **Ontology Merging**
  - Implementation: Java, OWL API
  - String-match
SPEED Simulator

Log File

RP45 is now connecting...
RP45 is now a Integration Peer with out semantic neighbors
Semantic Index: <<Cluster: 45>>

Simulation time: 1161 seconds
External indices: RandIndex=0.942 JaccardCoefficiet=0.646 FMIndex=0.785 Hubbert=0.752

Ontology-based Clustering in a Peer Data Management System
Experiments

• #Requesting Peers: 45
• Search Strategy
  – allClusters vs. limitedClusters
• allClusters
  – Semantic index is discarded
• limitedClusters
  – Multiple executions
  – Different orders of requesting peers
Experiments: limitedClusters

External Indices
“Golden Standard”
Hierarchical clustering Algorithm

Internal Indices
Similarity between peers of the same cluster
Dissimilarity between peers of the distinct cluster

Ontology-based Clustering in a Peer Data Management System
limitedClusters vs. allClusters

Search Strategy
External Indices
Cluster Threshold = 0.35

#Executions of SemMatch
External Indices
Cluster Threshold = 0.35

### External Indices

<table>
<thead>
<tr>
<th>Search Strategy</th>
<th>Rand Index</th>
<th>Jaccard Coefficient</th>
<th>Fowlkes-Mallows Index</th>
<th>Hubert's Statistic</th>
</tr>
</thead>
<tbody>
<tr>
<td>allClusters</td>
<td>0.970</td>
<td>0.675</td>
<td>0.794</td>
<td>0.778</td>
</tr>
<tr>
<td>limitedClusters</td>
<td>0.942</td>
<td>0.649</td>
<td>0.788</td>
<td>0.755</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>External Indices</th>
<th>#Executions of SemMatch</th>
</tr>
</thead>
<tbody>
<tr>
<td>SemMatch (CLO,CLO)</td>
<td>allClusters: 122, limitedClusters: 91</td>
</tr>
<tr>
<td>SemMatch (CLO,LO)</td>
<td>allClusters: 212, limitedClusters: 155</td>
</tr>
<tr>
<td>SemMatch (OS,LO)</td>
<td>allClusters: 0, limitedClusters: 271</td>
</tr>
</tbody>
</table>

LO = Local Ontology  CLO = Cluster Ontology  OS = Ontology Summary

Ontology-based Clustering in a Peer Data Management System

Universidade Federal de Pernambuco
Experiments

Transmitted Messages
External Indices
limitedClusters vs. allClusters
Cluster Threshold = 0.35

Query Routing
limitedClusters
Cluster Threshold = 0.35

Ontology-based Clustering in a Peer Data Management System
Conclusions

- Incremental process to cluster semantically similar peers in a PDMS
- Peers are organized according to a mixed P2P topology
- Ontologies are used to represent exported schemas
- Peer clustering is assisted by
  - Ontology matching
  - Ontology summarization
Contributions

- Ontology-based PDMS Architecture
- Ontology Matching Process
  - Determine most similar cluster
  - Determine cluster neighborhood
  - Search in the semantic index
- Ontology Summarization Process
  - Cluster Ontologies
- Incremental Peer Clustering Process
Future Work

• **SemMatch**
  – Consider properties of the concepts

• **OWLSum**
  – Apply transitivity rules in order to eliminate non-relevant concepts

• **Peer Clustering**
  – Improve semantic index, e.g. organization and search
  – Consider peer disconnection

• **Load Balancing**
  – Merging and split of clusters
Ontology-based Clustering in a Peer Data Management System

Ph.D. Thesis

Carlos Eduardo Santos Pires (cesp@cin.ufpe.br)

Advisor: Ana Carolina Salgado (acs@cin.ufpe.br)

April 27, 2009 - Recife, PE, Brazil