Distributed Spectral Cluster Management: A Method For Building Dynamic Publish/Subscribe Systems

Adnan Tariq, Boris Koldehofe, Gerald G. Koch and Kurt Rothermel

IPVS – Distributed Systems, Universität Stuttgart, Germany
Outline

• Introduction
• Contributions
• Subscription Clustering Problem
• Centralized Approach
• Distributed Approach
  ◦ Dimensionality Embedding
  ◦ K-mean Algorithm
• Evaluations and Results
• Conclusion
Introduction

- Publish/Subscribe is characterized by three aspects
  - Many to many communication
  - Selective dissemination of information
  - Decoupled interaction

Topic based
Information divided into topics

Content based
Restrictions on message content
\{ Stock = any ^ Price < 100 \}

Traditional System

P2P based systems
Reduce false positives
Subscription Clustering

Clustering requirements
• Reduction in false positives

Clustering criteria
• Structural similarities
• Event traffic based similarities

Limitations of existing systems
• Structural similarities  [Sub-2-Sub. 2006, Anceaume et al. 2006, DR-tree. 2010]
Our contributions

• Formulate subscription clustering problem
  ◦ Preserves expressiveness and use event traffic based similarity

• Identify centralized spectral methods to solve the clustering problem
  ◦ Effective in comparison to related approaches in P2P

• Propose methods to perform spectral clustering in distributed manner
  ◦ Accuracy closely matching that of centralized methods
  ◦ Significant reduction in computational time

• Develop an approach to maintain clusters in P2P environment
Subscription Clustering Problem

- Given **event traffic based similarities** between subscriptions maintain „k“ clusters such that
  - Overlap between event in different clusters is minimized
  - Clusters are balanced to ensure even event dissemination load

  ![Similarity graph of subscriptions](image)

  \[
  \text{minimize} \sum_{i=1}^{k} \frac{W(C_i, \overline{C_i})}{d_{C_i}}
  \]

  \[
  \text{maximize} \sum_{i=1}^{k} \frac{W(C_i, C_i)}{|C_i|}
  \]

- Both functions are **NP-hard in discrete domain**
  - Spectral analysis of similarity graph
Centralized Spectral Clustering

- Calculate similarities between subscriptions

\[ \text{sim} (S_i, S_j) = \frac{\text{Events matched by both subscriptions}}{\text{Total events received}} \]

1 → Consume same events, 0 → Completely disjoint

Similarity Matrix

<table>
<thead>
<tr>
<th></th>
<th>S_1</th>
<th>S_2</th>
<th>S_3</th>
<th>S_4</th>
<th>S_5</th>
</tr>
</thead>
<tbody>
<tr>
<td>S_1</td>
<td>0.3</td>
<td>1.0</td>
<td>0.2</td>
<td>0.8</td>
<td></td>
</tr>
<tr>
<td>S_2</td>
<td>0.3</td>
<td>0.2</td>
<td>1.0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>S_3</td>
<td>0.2</td>
<td>1.0</td>
<td>1.0</td>
<td>0.7</td>
<td></td>
</tr>
<tr>
<td>S_4</td>
<td>0.2</td>
<td>1.0</td>
<td>0.3</td>
<td>0.9</td>
<td></td>
</tr>
<tr>
<td>S_5</td>
<td>0.8</td>
<td>1.0</td>
<td>0.7</td>
<td>0.9</td>
<td></td>
</tr>
</tbody>
</table>
Centralized Spectral Clustering

• Calculate similarities between subscriptions

\[ \text{sim} (S_i, S_j) = \frac{\text{Events matched by both subscriptions}}{\text{Total events received}} \]

1 → Consume same events, 0 → Completely disjoint

• Convert similarities to Laplacian matrix

• Calculate „k“ eigenvectors with smallest eigenvalues
  ○ Embed subscriptions in „k“ dimensional space

• Perform k-mean algorithm
Performance of Centralized Clustering

- Drawbacks of centralized clustering
  - Bandwidth utilization: Periodic fetching of subscriptions and event histories
  - High computational time

Overlap → Wang et al. 2002
kmean → Casalicchio et al. 2007
Distributed Spectral Clustering (1)

- Dimensionality embedding
  - Calculation of eigenvectors
  - Similarities between all pairs of subscriptions

- k-mean algorithm
  - Iterative method
  - Distance of each point to the cluster centers
  - Selection of initial cluster centers

Assign \( i_1 \) to the cluster with nearest center \( w_j \) i.e., \(|i_1 - w_j^*| < |i_1 - w_j|\)
Distributed Spectral Clustering (2)

- Hierarchical grouping of peers
  - Each peer maintains a set of subscriptions \((S_1 \ldots S_m)\)
  - Group management depends on the no. of maintained subscriptions
Distributed Spectral Clustering (3)

- Each group maintains subset of subscriptions
  - Effects computation time and memory requirements
- Selection of Landmark subscriptions
  - Uniformly from the subscriptions maintained by the lower level group

![Diagram showing distribution of landmarks and subscriptions](image)
Dimensionality Embedding

- **Local embedding**
  - Separate embeddings of hierarchical groups ($X_L$)

- **Global embedding**
  - Root subscriptions defines global basis ($Y_G$)
  - Local embeddings are fixed according to global basis
  - Projection matrix ($H$) is derived such that embedding error is minimized

\[ Y_{L_1} = H_G X_{L_1} \]

\[ Y_{L_0} = H_{L_1} X_{L_0} \]
k-mean Algorithm

- Sampling Method
  - Cluster global landmarks to obtain centers of global clusters
k-mean Algorithm

- **Sampling Method**
  - Cluster global landmarks to obtain centers of global clusters

- **Hierarchical Method**
  - Selection phase – Identification of good cluster centers
  - Merge phase – Calculation of final clusters
Results: Scalability of Distributed Spectral Clustering

Computational time in comparison to centralized spectral clustering

Matrix Algebra: JAMA library
CPU: 2.7 GHz Intel core i7
RAM: 4GB
Results: Scalability of Distributed Spectral Clustering

Overall clustering time w.r.t. no. of subscriptions

Communication delay: [224ms, 384ms]
Percentage of landmarks: 25%
Levels of hierarchy: 4
Clustering method: Hierarchical
Results: Reduction in Message Overhead

\[ \text{Notification Cost} = \frac{\text{Overall traffic in the system}}{\text{No. of subscriptions matched by the events}} \]

Readings : 50 events
Levels of hierarchy : 2
No. of clusters : 25
Conclusion and Future work

• Conclusion
  ◦ An approach to perform spectral clustering in distributed manner
    ▪ Drastically reduce the time to perform clustering (86% - 99%)
    ▪ Effectively identify good quality clusters
    ▪ Significantly reduce the cost of event dissemination
  ◦ Propose methods are general and can be applied to other areas
    ▪ Document clustering, data mining

• Future work
  ◦ More sophisticated method for the selection of landmarks
    ▪ Maximum independent set
Questions?

Thank you for your attention!