On-Demand Cluster Analysis for Product Line Functional Requirements

Nan Niu, Steve Easterbrook

Department of Computer Science
University of Toronto, Canada
Email: {nn, sme}@cs.toronto.edu

Sept. 9, 2008
Core Assets: Benefit or Burden?

⇒ Reuse is planned, enabled, and enforced
  ⇒ Core assets: design for reuse
  ⇒ Products: design with reuse

⇒ Core assets ≠ actual product
  ⇒ Not worth building by itself
SPL Requirements Methods

» Based on domain analysis
  » Feature-Oriented Domain Analysis [Kang et al., 1990]
  » Product-line Requirements Specification [Faulk, 2001]

» Problems
  » domain dependence
  » heavyweight

problems
Requirements Clustering

⇒ Unsupervised learning
  ⇒ Organize and find structures in requirements
  ⇒ No manually-labeled training set
  ⇒ Amenable to automation

⇒ Essentials
  ⇒ Before clustering
    ⇒ Objects & attributes
  ⇒ After clustering
    ⇒ Name, summarize, display, explain
    ⇒ On-demand: stakeholders have different goals when performing cluster analysis
Functional Requirements Profiles

⇒ Why?
  ⇒ Salient features directly observable by all stakeholders
  ⇒ Basis for aligning and optimizing quality requirements

⇒ Extracting & modeling FRPs [RE’08]
  ⇒ Action-oriented concerns in the domain
  ⇒ Model user-visible system functionalities
  ⇒ Represented as “verb-direct object” pairs
FRPs & Semantic Cases

(a) Requirements list for the library MIS

| R₁ | A borrower can reserve a book. |
| R₂ | A borrower may cancel a reservation for a book. |
| R₃ | A librarian shall add a book to the library. |
| R₄ | The system shall notify the borrower when the reservation is available. |
| R₅ | Search what books are currently checked out by a borrower. |
| R₆ | Search all the reservations made by a borrower. |
| R₇ | The administrator shall add a user to the system. |

(b) FRPs

| FRP₁ | reserve book |
| FRP₂ | cancel reservation |
| FRP₃ | add book |
| FRP₄ | notify borrower |
| FRP₅ | search checkout |
| FRP₆ | check out book |
| FRP₇ | search reservation |
| FRP₈ | add borrower |
| FRP₉ | add librarian |

FRP₄: notify borrower

**Agentive:** system-to-be

**Objective:** student, professor, staff, visitor

**Conditional:** e-mail address

FRP₅: search checkout

**Agentive:** borrower, librarian

**Objective:** overdue, admissible

**Conditional:** identify borrower, check authentication
Outline

- Why, what, & how to cluster
- On-demand cluster analysis
- Concluding remarks
User-Centered Clustering

- **Stakeholder goals**
  - External utility; NOT internal implem. structure
  - Identify, browse, and prioritize features

- **Advance of literature**
  - Overlapping clusters to uncover crosscutting concerns
Overlapping Partitioning Cluster

ɐ Goals

ɐ Maximizing the average number of objects in a cluster
ɐ Maximizing the distances among cluster center objects

ɐ Do not require a pre-fixed number of clusters
ɐ Allow the user to explore features with different granularities

ɐ Threshold is set by domain knowledge
Example

(a)

<table>
<thead>
<tr>
<th>A_1 : system-to-be</th>
<th>FRP_1</th>
<th>FRP_2</th>
<th>FRP_3</th>
<th>FRP_4</th>
<th>FRP_5</th>
<th>FRP_6</th>
<th>FRP_7</th>
<th>FRP_8</th>
<th>FRP_9</th>
<th>FRP_{10}</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>A_2 : borrower</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>A_3 : librarian</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>A_4 : administrator</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>A_5 : book</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>A_6 : reservation</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>A_7 : checkout</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>A_8 : e-mail address</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>A_9 : ID &amp; password</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

(b)

\[
\begin{array}{cccccccccc}
\text{FRP}_1 & \text{FRP}_2 & \text{FRP}_3 & \text{FRP}_4 & \text{FRP}_5 & \text{FRP}_6 & \text{FRP}_7 & \text{FRP}_8 & \text{FRP}_9 & \text{FRP}_{10} \\
\text{FRP}_2 & 1.00 & 0.74 & 0.38 & 0.17 & 0.30 & 0.38 & 0.58 & 0.30 & 0.10 & 0.30 \\
\text{FRP}_3 & 1.00 & 0.17 & 0.17 & 0.38 & 0.17 & 0.74 & 0.38 & 0.14 & 0.38 & 0.38 \\
\text{FRP}_4 & 1.00 & 0.00 & 0.38 & 0.17 & 0.38 & 0.14 & 0.38 & 0.14 & 0.38 & 0.38 \\
\text{FRP}_5 & 1.00 & 0.14 & 0.17 & 0.14 & 0.38 & 0.14 & 0.38 & 0.14 & 0.38 & 0.38 \\
\text{FRP}_6 & 1.00 & 0.38 & 0.58 & 0.30 & 0.30 & 0.58 & 0.30 & 0.30 & 0.58 & 0.30 \\
\text{FRP}_7 & 1.00 & 0.14 & 0.14 & 0.00 & 0.14 & 0.14 & 0.00 & 0.14 & 0.14 & 0.00 \\
\text{FRP}_8 & 1.00 & 0.30 & 0.30 & 0.14 & 0.30 & 0.30 & 0.14 & 0.30 & 0.30 & 0.14 \\
\text{FRP}_9 & 1.00 & 0.58 & 0.58 & 0.58 & 0.58 & 0.58 & 0.58 & 0.58 & 0.58 & 0.58 \\
\text{FRP}_{10} & 1.00 &       &       &       &       &       &       &       &       &       \\
\end{array}
\]

(c)

\[
C_1 = \{ \text{FRP}_1, \text{FRP}_2, \text{FRP}_7 \} \\
C_2 = \{ \text{FRP}_5, \text{FRP}_7, \text{FRP}_{10} \} \\
C_3 = \{ \text{FRP}_8, \text{FRP}_9, \text{FRP}_{10} \}
\]
Example’s Result

- $C_1$: reservation
- $C_2$: searching
- $C_3$: admin

- $FRP_1$: reserve book
- $FRP_2$: cancel reservation
- $FRP_4$: notify borrower
- $FRP_5$: search checkout
- $FRP_6$: check out book
- $FRP_7$: search reservation
- $FRP_8$: add borrower
- $FRP_9$: add librarian
- $FRP_{10}$: check authentication
Outline

Why, what, & how to cluster

On-demand cluster analysis
  - User-centered clustering
  - Design-driven clustering

Concluding remarks
Design-Driven Clustering

⇒ Stakeholder goals
  ⇒ System decomposition & modularization
  ⇒ CCCs; the rest shouldn’t overlap

⇒ Advance of literature
  ⇒ Information-theoretic clustering: minimize the information loss during clustering


Information Theory Preliminaries

⇒ Entropy

⇒ A measure of disorder
⇒ A={FRP₁, FRP₂, ..., FRPₙ}
⇒ \( H(A) = - \sum_{a_i} p(a_i) \log_2 p(a_i) \)

⇒ Conditional entropy

⇒ A measure of uncertainty with which we can predict the value of B given that a value of A appears
⇒ B={Attribute₁, Attribute₂, ..., Attributeₘ}
⇒ \( H(B|A) = - \sum_{a_i} p(a_i) \sum_{b_j} p(b_j|a_i) \log_2 p(b_j|a_i) \)
## Example

<table>
<thead>
<tr>
<th></th>
<th>FRP₁</th>
<th>FRP₂</th>
<th>FRP₃</th>
<th>FRP₄</th>
<th>FRP₅</th>
<th>FRP₆</th>
<th>FRP₇</th>
<th>FRP₈</th>
<th>FRP₉</th>
<th>FRP₁₀</th>
</tr>
</thead>
<tbody>
<tr>
<td>A₁</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1/3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>A₂</td>
<td>1/4</td>
<td>1/3</td>
<td>0</td>
<td>1/3</td>
<td>1/4</td>
<td>1/3</td>
<td>1/4</td>
<td>1/4</td>
<td>0</td>
<td>1/4</td>
</tr>
<tr>
<td>A₃</td>
<td>0</td>
<td>0</td>
<td>1/3</td>
<td>0</td>
<td>1/4</td>
<td>0</td>
<td>1/4</td>
<td>0</td>
<td>1/4</td>
<td>1/4</td>
</tr>
<tr>
<td>A₄</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1/4</td>
<td>1/4</td>
<td>1/4</td>
</tr>
<tr>
<td>A₅</td>
<td>1/4</td>
<td>0</td>
<td>1/3</td>
<td>0</td>
<td>0</td>
<td>1/3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>A₆</td>
<td>1/4</td>
<td>1/3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1/4</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>A₇</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1/4</td>
<td>1/3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>A₈</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1/3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1/4</td>
<td>1/4</td>
<td>0</td>
</tr>
<tr>
<td>A₉</td>
<td>1/4</td>
<td>1/3</td>
<td>1/3</td>
<td>0</td>
<td>1/4</td>
<td>0</td>
<td>1/4</td>
<td>1/4</td>
<td>1/4</td>
<td>1/4</td>
</tr>
</tbody>
</table>

Each column stores the distribution $p(B|A=a_i)$.
Mutual Information

\[ I(B;C) = H(B) - H(B|C) \]

\( C = \{ \text{Cluster}_1, \text{Cluster}_2, \ldots, \text{Cluster}_k \} \)

Measures the amount of information that two variables hold about each other

Quantifies the information about the value of B (domain attributes) provided by the identity of an FRP-cluster (a given value of C)

Goal: maximize the value of \( I(B;C) \)
Example

<table>
<thead>
<tr>
<th></th>
<th>FRP_1</th>
<th>FRP_2</th>
<th>FRP_3</th>
<th>FRP_4</th>
<th>FRP_5</th>
<th>FRP_6</th>
<th>FRP_7</th>
<th>FRP_8</th>
<th>FRP_9</th>
<th>FRP_10</th>
</tr>
</thead>
<tbody>
<tr>
<td>FRP_1</td>
<td>-</td>
<td>0.03</td>
<td>0.09</td>
<td>0.14</td>
<td>0.10</td>
<td>0.09</td>
<td>0.05</td>
<td>0.10</td>
<td>0.15</td>
<td>0.10</td>
</tr>
<tr>
<td>FRP_2</td>
<td>-</td>
<td>0.13</td>
<td>0.13</td>
<td>0.09</td>
<td>0.13</td>
<td>0.03</td>
<td>0.09</td>
<td>0.14</td>
<td>0.09</td>
<td></td>
</tr>
<tr>
<td>FRP_3</td>
<td>-</td>
<td></td>
<td>0.20</td>
<td>0.09</td>
<td>0.13</td>
<td>0.09</td>
<td>0.14</td>
<td>0.09</td>
<td>0.09</td>
<td></td>
</tr>
<tr>
<td>FRP_4</td>
<td>-</td>
<td></td>
<td></td>
<td>0.14</td>
<td>0.13</td>
<td>0.14</td>
<td>0.09</td>
<td>0.14</td>
<td>0.14</td>
<td></td>
</tr>
<tr>
<td>FRP_5</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td>0.09</td>
<td>0.05</td>
<td>0.10</td>
<td>0.10</td>
<td>0.05</td>
<td></td>
</tr>
<tr>
<td>FRP_6</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.14</td>
<td>0.14</td>
<td>0.20</td>
<td>0.14</td>
<td></td>
</tr>
<tr>
<td>FRP_7</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.10</td>
<td>0.10</td>
<td>0.05</td>
<td></td>
</tr>
<tr>
<td>FRP_8</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.05</td>
<td>0.05</td>
<td></td>
</tr>
<tr>
<td>FRP_9</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.05</td>
<td></td>
</tr>
<tr>
<td>FRP_10</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Complexity: $O(n^2 \log n)$
Example's Result
A Fuller Example
Experiences

⇒ On-demand & coherent framework
   ⇒ Persistence: early aspect module in data-dominant domains

⇒ Domain knowledge
   ⇒ Threshold
   ⇒ Variability, e.g., mandatory or optional
   ⇒ Name & explain the cluster

⇒ Quality requirements
   ⇒ FRPs help align NFRs & explore high-level features [Niu & Easterbrook, IEEE Software, 2007]
Concluding Remarks

 ⇒ Current work
   ⇒ Clustering provides automatic support for domain analysis
   ⇒ On-demand recognizes stakeholders’ different goals: user vs. designer
   ⇒ Advanced literature: overlapping clusters & minimal information loss

 ⇒ Future work
   ⇒ Intermediate clustering to address scalability
   ⇒ Semantic attributes & relationships
   ⇒ Weighting schemes