KOD Team - Nantes France

Post-Processing of Discovered Association Rules Using Ontologies

Claudia Marinica
Fabrice Guillet and Henri Briand

The Second International Workshop on Domain Driven Data Mining (DDDM 2008)
IEEE International Conference on Data Mining (ICDM 2008)

15/12/2008
Outline

1. Research Field & Problem Definition
2. Related Work
3. Proposed Approach
4. Conclusions and Future Works
Research Field & Problem Definition

- **Association rule mining:**
  - Huge volume of discovered association rules:
    - Interestingness ? : Not-actionable, already known
    - difficult to analyze manually
  
  => help the user with an efficient reducing rule number

- **Goal:**
  - Efficient post-processing task integrating user knowledge

- **Approach :**
  - GI (Rule Schemas) + Ontologies
  
  => supervising association rule mining
Outline

1. Research Field & Problem Definition
2. Related Work
3. Proposed Approach
4. Conclusions and Future Works
Related work

- **Actionability - Interestingness measures** *(Silberschatz et Tuzhilin, 1996)*:
  - Numerical indicators defining association rule quality

- **Objective measures** *(Piatetsky-Shapiro, 1991, Guillet and Hamilton, 2007)*
  - Statistical indicators evaluating the strength of a rule over data
  - Limitation: huge number of discovered rules and almost of them are not interesting for the user

- **Subjective measures** *(Liu et al., 1999, Padmanabhan et Tuzhilin, 1997)*
  - Depend on user goals, beliefs and expectations
  - Combined with supervised algorithms in order to select interesting rules only
Related work

- **Rule Templates** (Klemettinen et al., 1994)
  - Model based on templates

- **Belief-based system** (Silberschatz et Tuzhilin, 1995)
  - Hard Beliefs and Soft Beliefs

- **Fuzzy matching technique** (Liu et Hsu, 1996)
  - User beliefs presented as fuzzy rules

- **Logical contradiction** (Padmanabhan et Tuzhilin., 1997)

- **User Knowledge Classification** (Liu et al., 1999)
  - General Impression (GI),
  - Reasonably Precise Concept (RPC), Precise Knowledge (PK)
Related work

- General Impressions (GI) / Reasonably Precise Concepts (RPC)

  - $gi/rpc \ (< S_1, S_2, \ldots (\rightarrow) \ldots S_n >)$
    - $S_i$ – elements of a taxonomy of items
      - expressions: *, +, ?

  - Example:
    - $gi (\ < \{\text{milk, cheese}\}^*; \text{Fruit}+, \text{beef} >)$
    - $rpc (\ < \text{Meat, Dairy\_product} \rightarrow \{\text{grape, apple}+\} >)$

  ![Diagram showing a taxonomy of food items](image)

  [Liu et al., 1999]
Ontology - an explicit specification of a conceptualization. A conceptualization is an abstract, simplified view of the world that we wish to represent for some purpose (Gruber T.R., 1993).

- Set of concepts \( C = \{ C_1, C_2, \ldots, C_o \} \)
- Set of relations / properties \( R = \{ R_1, R_2, \ldots, R_r \} \)
- Subsuming relation, \( \leq \):
  - \( C_1 \leq C_2 \) – if the concept \( C_1 \) subsumes the concept \( C_2 \).

Goal: organizing database items in an hierarchical structure
Related work

- Domain ontologies

- First propositions using ontologies:
  - **Generalized Association Rules** (Srikant & Agrawal, 1995)
    - Item hierarchy using taxonomies
  - **Raising** (Chen et al., 2003, Zhou and Geller, 2007)
    - Generalizing rules improving support and keeping the same confidence level (taxonomy)
1. Research Field & Problem Definition
2. Related Work
3. Proposed Approach
4. Conclusions and Future Works
Proposed Approach

- Starting points:
  - Generalized association rules -> **ontology**
  - General Impressions -> **Rule Schemas**

- Improving post-processing phase:
  - Applying **operators** over rules in order to reduce rule number
**Proposed Approach**

- **Example:**

\[
\text{gi (< \{milk, cheese\}*; Fruit+, beef >)}
\]

- **Define new concepts as:**
  - BioProducts
  - DietProducts

Using an ontology

\[
\text{gi (< BioProducts, DietProducts >)}
\]

or

\[
\text{rpc (< DietProducts \rightarrow BioProducts >)}
\]
Proposed Approach

KOD Team

Ontology

Rule Schemas

Operators

User Knowledge

DB

Association Rules

Post-processing

Filtered rules

Association Rules Mining

Post-processing step
Proposed Approach

- Questionnaire database, about client satisfaction concerning accommodation (Nantes Habitat)

- Annual study (since 2003) on 1500 out of a total of 5000 clients => 9000 transactions

- 67 questions with 6 possible answers:
  - “quite”=1, ”rather”=2, “rather not”=3, “not at all”=4
  - non applicable cases - 95/96
  - the client doesn’t know the answer – 99

- Example:
  - item q1=1 => question q1=“Is your district transport practical?” with the answer 1=“quite”.

Claudia Marinica
15/12/2008
Proposed Approach

- Developing an item ontology with the validation of the domain expert
- Revising the ontology

- Concept types:
  - Leaf-concepts
  - Generalized concepts
  - Defined concepts

- Connection with the database

SatisfactionConfortApartement →
{q44=1, ..., q48=1, q44=2, ..., q48=2}
Proposed Approach

- Extend General Impressions with ontologies

- Syntax: $gi/rpc (< S_1, S_2, \ldots (\rightarrow) \ldots S_n >)$
  - $S_i$ – element of the ontology

- Example:
  - $RS: (< \text{UnsatComfortApartment} \rightarrow \text{UnsatHostListen} >)$

- Operators – applied over rule schemas
  - Pruning $P(RS)$
  - Filtering
    - Conforming $C(RS)$
    - Unexpectedness $U(RS)$
Proposed Approach

User Knowledge
- Results -

- Extract the association rules (Weka software):
  - min support = 2% and max support = 30%
  - min confidence = 80%

=> 82,159 association rules.

<table>
<thead>
<tr>
<th>Rule Schema</th>
<th>Operator</th>
<th>Rules</th>
</tr>
</thead>
<tbody>
<tr>
<td>RS₁: &lt;SatFirstAppearance → BuildingsCondition&gt;</td>
<td>P(RS₁)</td>
<td>1,974</td>
</tr>
<tr>
<td>RS₂: &lt;SatGarbagePlace → SatCommonPlace&gt;</td>
<td>P(RS₂)</td>
<td>8,743</td>
</tr>
<tr>
<td>RS₃: &lt;UnsatPrice, UnsatCalmDistrict&gt;</td>
<td>C(RS₃)</td>
<td>7</td>
</tr>
<tr>
<td>RS₄: &lt;SatComfortApartment → SatDelais&gt;</td>
<td>C(RS₄)</td>
<td>1,024</td>
</tr>
<tr>
<td>RS₅: &lt;UnsatComfortApartment → UnsatAgencyReceiving&gt;</td>
<td>U(RS₅)</td>
<td>4</td>
</tr>
</tbody>
</table>
Proposed Approach

<table>
<thead>
<tr>
<th>Antecedent</th>
<th>Consequent</th>
<th>Confidence</th>
<th>Support</th>
</tr>
</thead>
<tbody>
<tr>
<td>q62=4,q64=4</td>
<td>q63=4</td>
<td>0.852</td>
<td>0.019</td>
</tr>
<tr>
<td>q64=4,q97=4</td>
<td>q73=4</td>
<td>0.805</td>
<td>0.019</td>
</tr>
<tr>
<td>q62=4,q72=4</td>
<td>q63=4</td>
<td>0.815</td>
<td>0.020</td>
</tr>
<tr>
<td>q58=4,q59=4,q62=4</td>
<td>q63=4</td>
<td>0.815</td>
<td>0.019</td>
</tr>
</tbody>
</table>

Results for:

*RS5*: (< UnsatComfortApartment $\rightarrow$ UnsatAgencyReceving >)

*Operator*: Unexpectedness
Outline

1. Research Field & Problem Definition
2. Related Work
3. Proposed Approach
4. Conclusions and Future Works
Conclusions and Future Work

- Proposed a new approach of post-processing association rules integrating user knowledge
  - GI (Rule Schemas) + Ontologies
  - \(\Rightarrow\) supervising association rules mining

- Testing our proposition on a real life study case and working in a collaboration with an expert (Nantes Habitat)

- New results and new filters integrated in the tool

15/12/2008
Claudia Marinica
We would like to thank *Nantes Habitat*, the Public Housing Unit in Nantes, France, and more specially Ms. Christelle Le Bouter, for supporting our work.