Relevance Feedback for Association Rules by Leveraging Concepts from Information Retrieval

Georg Ruß, russ@iws.cs.uni-magdeburg.de
Mirko Böttcher, mail@mirkoboettcher.de
Prof. Dr. Rudolf Kruse, kruse@iws.cs.uni-magdeburg.de

December 12th, 2007

Outline

Motivation

- Large amounts of transactional data
- Association rule mining yields rules as a condensed representation
- Form: IF $item_1$, $item_2$, $\ldots$, $item_n$ THEN item_m
- Problem: still too many rules to analyze
- Topic: find interesting association rules

Association Rules – Formalization

- Set $\mathcal{D}$ of *transactions* $\mathcal{T} \in \mathcal{D}$.
- Transaction $\mathcal{T}$ is a subset of a set of items $\mathcal{L}$.
- A subset $\mathcal{X} \subseteq \mathcal{L}$ is called *itemset*.
- A transaction $\mathcal{T}$ *supports* an itemset $\mathcal{X}$ if $\mathcal{X} \subseteq \mathcal{T}$.
- An association rule r is an expression $\mathcal{X} \rightarrow \mathcal{Y}$ where $\mathcal{X}$ and $\mathcal{Y}$ are itemsets, $|\mathcal{Y}| > 0$ and $X \cap Y = \varnothing$.
  - $\mathcal{X}$: body, $\mathcal{Y}$: head
  - Rule reliability: *confidence* $\text{conf}(r) = P(\mathcal{Y} \mid \mathcal{X})$
  - Statistical significance: *support* $\text{supp}(r) = P(\mathcal{X}\mathcal{Y})$
  - Time series: confidence and support of one rule over time

Linking to Information Retrieval

- Interestingness of rules is subjective.
- Finding interesting rules requires user input.
- Manual specification of user's knowledge
  - key aspects are often forgotten
  - requires expert user
  - knowledge changes
  - hard to specify at beginning of analysis

Information Retrieval – Relevance Feedback

- Automatic acquisition of user's knowledge through actions
  - user rates what he sees
  - easy (binary) decision: interesting / not interesting
  - system collects user's choices and updates results
- *Relevance Feedback* known from Information Retrieval
  - association rules are presented (possibly pre-ordered)
  - user can examine and rate them
  - an internal ranking is adapted
  - best results are presented
  - cycle starts over

Rule Representation – Informal

- Use existing algorithms for relevance feedback from IR
- Represent rules as vectors
  - $$\vec{r} = (\overbrace{\underbrace{r_1, \ldots, r_b}_{\texttt{body}}, \underbrace{r_{b+1}, \ldots, r_{b+h}}_{\texttt{head}}}^{\texttt{symbolic}}, \underbrace{r_{b+h+1}, \ldots, r_{b+h+t}}_{\texttt{timeseries}}) \quad (1)$$
  - item weights: TF-IDF approach
    - high weight: term frequent in rule (TF), but less frequent in rule set (IDF)
    - filters commonly used terms, captures perceived relevance

Rule Representation – Maths

- term frequency

$$tf(x, r) = \begin{cases} 1 & \text{if } x \in r, \\ 0 & \text{otherwise.} \end{cases} \tag{2}$$

- inverse document frequency

$$idf(x, R) = 1 - \frac{\ln |r : r \in R \wedge x \in r|}{\ln |R|} \tag{3}$$

- A rule's feature vector is filled as follows:
  - $$r_i = tf(x_i, r) \cdot idf(x_i, R), \quad i = 1, \ldots, b \tag{4}$$
  - $$r_{b+j} = tf(x_j, r) \cdot idf(x_j, R), \quad j = 1, \ldots, h \tag{5}$$
  - $r_{timeseries}$ . . . respective time-variant properties of rule

Interestingness of Rules

- Idea: compare the same features of different rules
- Interestingness based on (dis-)similarity
- Six combinations deemed interesting:

similar \ *dissimilar*	head	body	time series	symbolic
head	-	ω_4	ω_5	-
body	ω_1	-	ω_6	-
time series	-	-	-	ω_2
symbolic	-	-	ω_3	-

Table: Interestingness Matrix

Pairwise Similarity

- Similarity between rules as measure of interestingness
- Similarity can easily be computed by similarity measures for vectors
- Cosine similarity:

$$sim(\vec{r}, \vec{s}) = \frac{\sum_{i=1}^{n} r_i s_i}{\sqrt{r_i^2}\sqrt{s_i^2}} \tag{6}$$

- Dissimilarity:

$$dissim(\vec{r}, \vec{s}) = 1 - sim(\vec{r}, \vec{s}) \tag{7}$$

Similarity Aggregation – first step

- Similarity between rule and rule set:

$$sim_{rs}(\vec{r}, R) = \Omega(\{sim(\vec{r}, \vec{s}_1), \ldots, sim(\vec{r}, \vec{s}_m)\}) \tag{8}$$

- Dissimilarity analogously to Equation 7:

$$dissim_{rs}(\vec{r}, R) = 1 - sim_{rs}(\vec{r}, R) \tag{9}$$

Similarity Aggregation – second step

- Use OWA operator to aggregate single similarities:
  - weighting vector $W = (w_1, w_2, \ldots, w_n)^T$ with $w_j \in [0, 1]$ and $\sum_{j=1}^{n} w_j = 1$
  - $$\Omega(\{s_1, s_2, \ldots, s_n\}) = \sum_{j=1}^{n} w_j b_j \quad , \tag{10}$$

    with b_j being the j-th largest of the s_i.

Relevance Scoring

- Score calculation for each association rule
- User selects rule r as interesting
  - determine interesting combinations:
  - rules with similar head, but different body
  - rules with similar body, but dissimilar head
  - six combinations (see Table 1)
- Calculate weighted sum of the score part in those six combinations
- Yields a relevance score for each association rule
- Sort rules by score – interesting ones assumed to have high score

Conclusion

- Similarity-based interestingness of association rules
- Incorporation of relevance feedback to find interesting rules
- User-specific, automatic adaptation
- Simple relevance scoring to assess interestingness