Relevance Feedback on Association Rules

Finding the most interesting rules

Georg Ruß

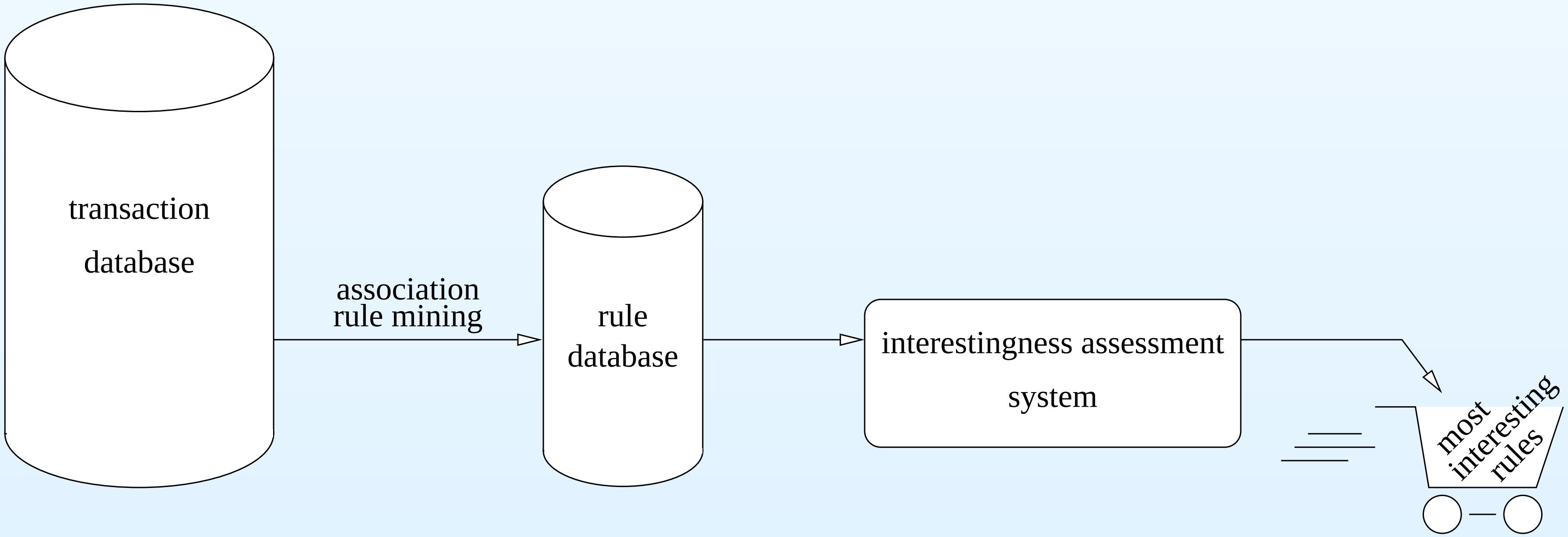
Otto-von-Guericke Universität Magdeburg

British Telecommunications plc

28 September 2006

Motivation

- Large relational databases containing e.g. survey data
- Database can be described by association rules
- Large number of association rules mined
- Find most interesting rules to support business decisions
- Take user's knowledge and preferences into account



Outline

1. Association Rules & Time Series
2. Interestingness / Preprocessing
3. Relevance Feedback
4. Results

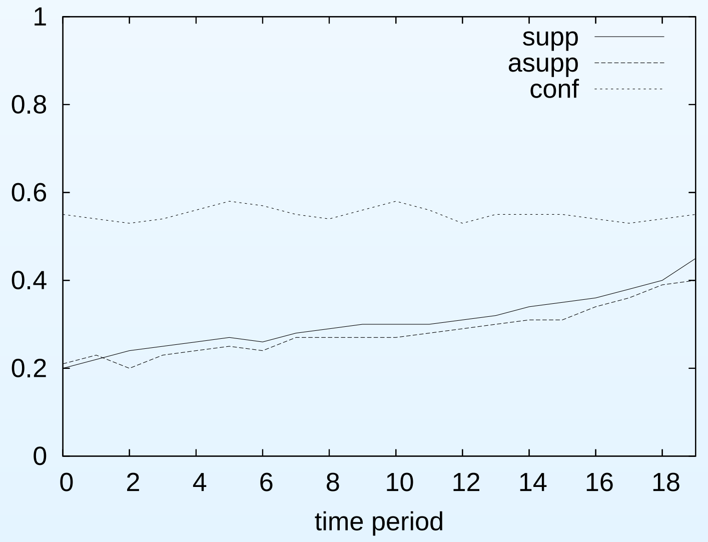
Association Rules - basic properties

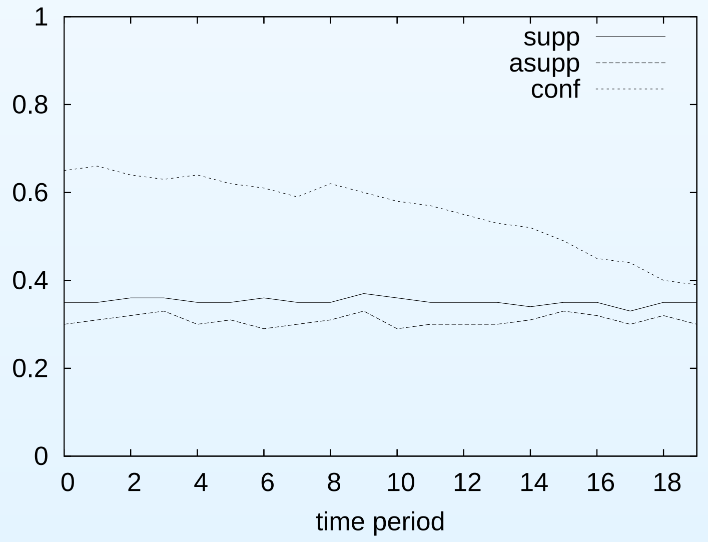
- rule symbolic features
  - transactions: $\mathcal{T} \in \mathcal{D}$
  - items: $\mathcal{L}$
  - association rule r: $\mathcal{X} \Rightarrow \mathcal{Y}$ with $\mathcal{X}$,$\mathcal{Y}$ sets of items
  - $\mathcal{X}$ “body”, $\mathcal{Y}$ “head”

- rule numeric features
  - $\mathrm{support}(\mathrm{r}) := \frac{|\{\mathcal{T} \in \mathcal{D} | \mathcal{X} \cup \mathcal{Y} \subseteq \mathcal{T}\}|}{|\mathcal{D}|}$
  - $\mathrm{confidence}(\mathrm{r}) := \frac{|\{\mathcal{T} \in \mathcal{D} | \mathcal{X} \cup \mathcal{Y} \subseteq \mathcal{T}\}|}{|\{\mathcal{T} \in \mathcal{D} | \mathcal{X} \subseteq \mathcal{T}\}|}$
  - $\mathrm{asupp}(\mathrm{r}) := \frac{|\{\mathcal{T} \in \mathcal{D} | \mathcal{X} \subseteq \mathcal{T}\}|}{|\mathcal{D}|}$

Association Rules - extended properties

- rules are mined from database snapshots
- numeric properties can be obtained from different snapshots
- generation of a time series of a rule for *conf, supp, asupp*





Association Rules - mined from customer survey data

- Items
  - `TEC={ADSL,CABLE,WIMAX}`,
  - `AGE={18-35,36-50,51-65,66+}`,
  - `SEX={M,F}`,
  - `WCC={YES,NO}`,
  - `SAT={VSAT,SAT,DISSAT}`
- Association rule: $\mathcal{X} \Rightarrow \mathcal{Y}$
- Example:
  - `TEC=ADSL,AGE=18-35`$\Rightarrow$`WCC=YES`
  - with an upward trend in confidence
- this type of association rule data was available for data analysis
  $\Rightarrow$ find most interesting rules from a large set (>8k)

Interestingness of rules - literature approaches

- most approaches based on symbolic part of rules only
  - contradiction, surprisingness, unexpectedness
  - usage of a “knowledge base” for above descriptors
    - hard to specify
    - requires expert
    - special specification languages
- often rules are discarded, but shouldn’t be
- time series information should be used

- $\Rightarrow$ develop a new interestingness definition, based on similarities between rules’ components

Interestingness of rules - link to information retrieval

- Querying - look for the most interesting rule
- Filtering - pick out the uninteresting rules
- Pattern Matching - find interesting rules that match a pattern
- Relevance Feedback - collect user's knowledge interactively
- Ranking - do not discard rules, but score and rank them

- $\Rightarrow$ Text Retrieval - consider association rules as documents?

Interestingness of rules = text retrieval?

- novel idea: consider association rules as text documents
- generate feature vectors as in text retrieval operations
- two-part feature vector:
  - part 1: symbolic, tf-idf weights of items
  - part 2: numeric, time series of rule

$$\vec{r} = (\underbrace{\overbrace{r_1, \ldots, r_b}^{\text{body}}, \overbrace{r_{b+1}, \ldots, r_{b+h}}^{\text{head}}}_{\text{symbolic}}, \underbrace{r_{b+h+1}, \ldots, r_{b+h+t}}_{\text{timeseries}})$$

Interestingness of rules - a similarity-based definition

- descriptors work between rule and knowledge base
  - surprising: if same antecedent, but different consequent
  - contradicting: if a related rule predicts the opposite
- similarity-based interestingness definition
- six interesting combinations of head, body, and time series

similar \ *dissimilar*	head	body	time series	head+body
head	-	+	+	-
body	+	-	+	-
time series	-	-	-	+
head+body	-	-	+	-

Interestingness of rules - examples

- similar antecedent but different consequent

	rule symbolic		*rule trend*	
#	body/antecedent	head/consequent	support	confidence
1	TEC=ADSL	WCC=YES	stable	stable
5	TEC=ADSL, AGE=36-50	WCC=NO	stable	stable

- similar rule - dissimilar time series

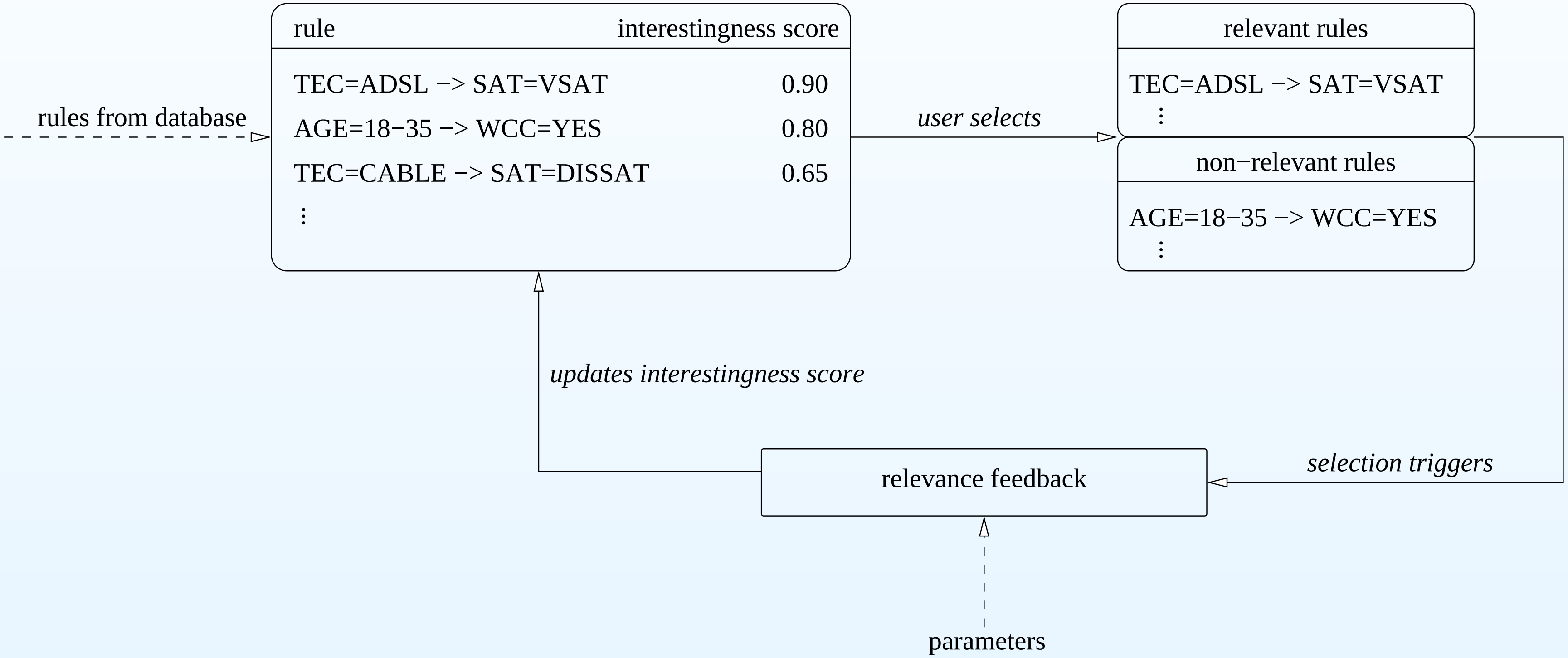
3	AGE=66+	VOL=NONE	down	down
7	AGE=66+	VOL=10GB	up	up

Relevance Feedback - introduction

Solves the problem of acquiring the "knowledge base".

- User implicitly has prior knowledge about the domain
- Interestingness of a rule depends on the user:
  - Which rules has the user seen before?
  - What is the user's search strategy?
- Collect the user's assumptions by having him select (non)relevant rules – construct the knowledge base
- Assign interestingness scores to the remaining rules.

Relevance Feedback - overview



Relevance Feedback - maths [1/5]

- user-selected sets of rules: R_{rel}, R_{nrel}
- A new interestingness score for each rule is calculated:

$$I(\vec{r}, R_{rel}, R_{nrel}, t) = \underbrace{w_{rel}\Phi(\vec{r}, R_{rel})}_{\text{relevant rules}} + \underbrace{w_{nrel}\Psi(\vec{r}, R_{nrel})}_{\text{non-relevant rules}}$$

- weights w_{rel} and w_{nrel} control the balance between the decisions of relevance and non-relevance

Relevance Feedback - maths [2/5]

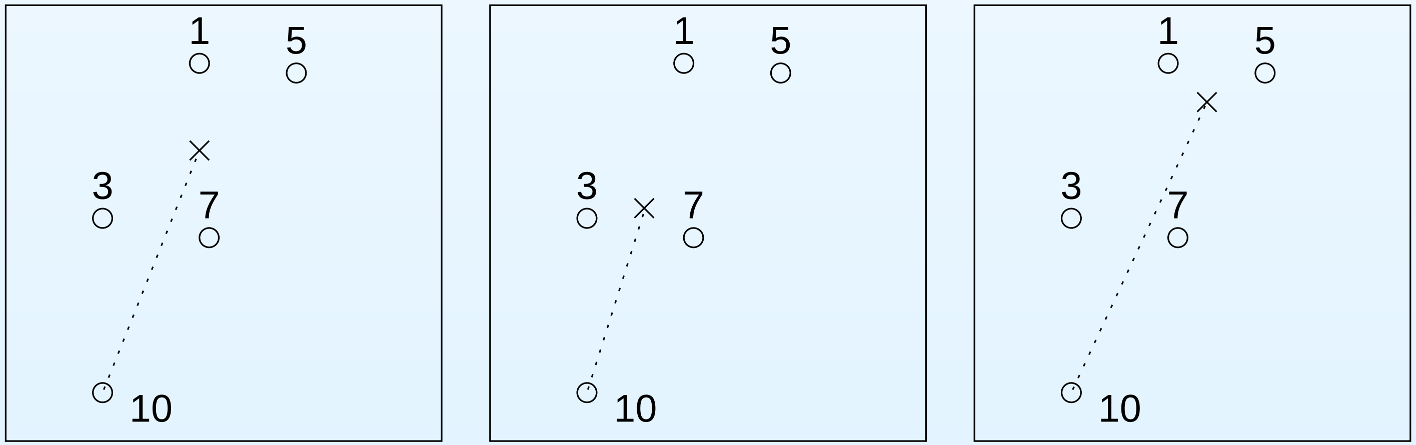
- remember the similarity-based interestingness definition

$$\begin{aligned}
\Phi(\vec{r}, R_{\text{rel}}) = {} & \omega_1 \cdot sim_{rs}(\vec{r}_{\text{body}}, R_{\text{rel}}) \cdot dissim_{rs}(\vec{r}_{\text{head}}, R_{\text{rel}}) \\
& + \omega_2 \cdot sim_{rs}(\vec{r}_{\text{time}}, R_{\text{rel}}) \cdot dissim_{rs}(\vec{r}_{\text{sym}}, R_{\text{rel}}) \\
& + \omega_3 \cdot sim_{rs}(\vec{r}_{\text{sym}}, R_{\text{rel}}) \cdot dissim_{rs}(\vec{r}_{\text{time}}, R_{\text{rel}}) \\
& + \omega_4 \cdot sim_{rs}(\vec{r}_{\text{head}}, R_{\text{rel}}) \cdot dissim_{rs}(\vec{r}_{\text{body}}, R_{\text{rel}}) \\
& + \omega_5 \cdot sim_{rs}(\vec{r}_{\text{head}}, R_{\text{rel}}) \cdot dissim_{rs}(\vec{r}_{\text{time}}, R_{\text{rel}}) \\
& + \omega_6 \cdot sim_{rs}(\vec{r}_{\text{body}}, R_{\text{rel}}) \cdot dissim_{rs}(\vec{r}_{\text{time}}, R_{\text{rel}})
\end{aligned}$$

- weights $w_1, \ldots, w_6$: allow to model the user’s search strategy

Relevance Feedback - maths [3/5]

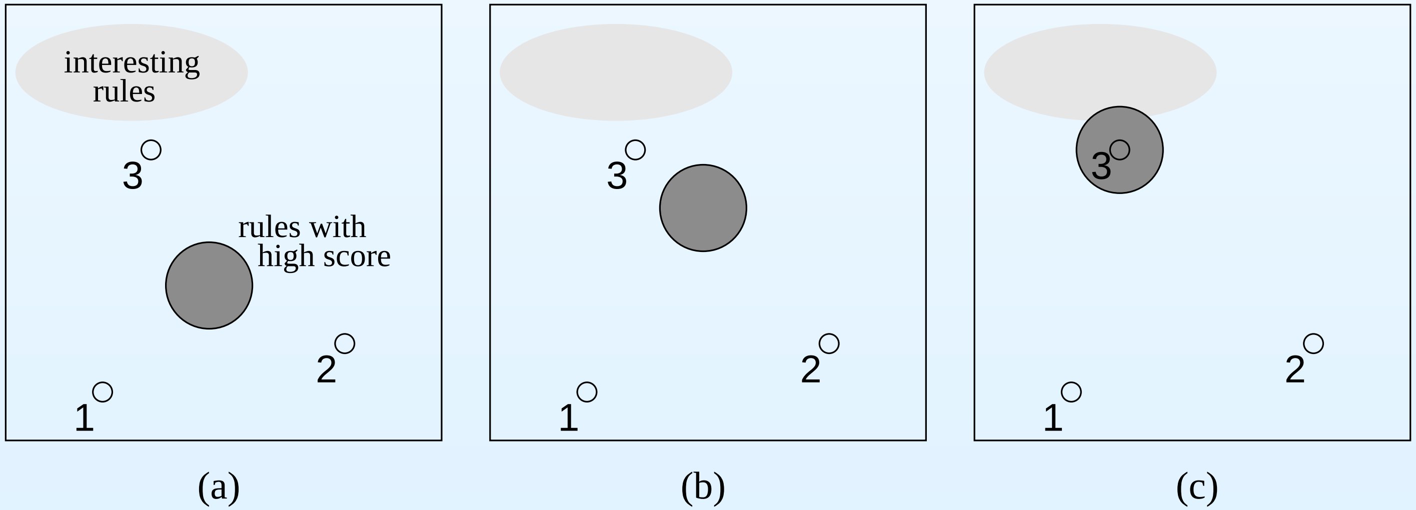
- Problem: define similarity between a vector and a set.
  - aggregation operator needed
  - OWA – Ordered Weighted Average operator from the Fuzzy domain
  - very flexible; can also imitate mean, max, min
  - allows to emphasise certain similarities (a)-mean, (b)-high (c)-low



(a) (b) (c)

Relevance Feedback - maths [4/5]

- Problem: model the user's importancy expectation.
  - "old rules" (from the first uninformed relevance choices) become less important as new ones are found
  - most recent rules are of higher importance to the relevance feedback
  - τ – decay of the importancy of a selected rule
  - most recent selection = highest influence on ranking (a) $\tau = 0$, (b) $\tau = 0.5$, (c) $\tau = 1$



Relevance Feedback - maths [5/5]

- Overall similarity calculation:

$$sim_{rs}(\vec{v}, R) \quad = \quad OWA(\{(\tau \cdot sim(\vec{v}, \vec{r_1})), \ldots, (\tau \cdot sim(\vec{v}, \vec{r_m}))\})$$

- dissimilarity calculation

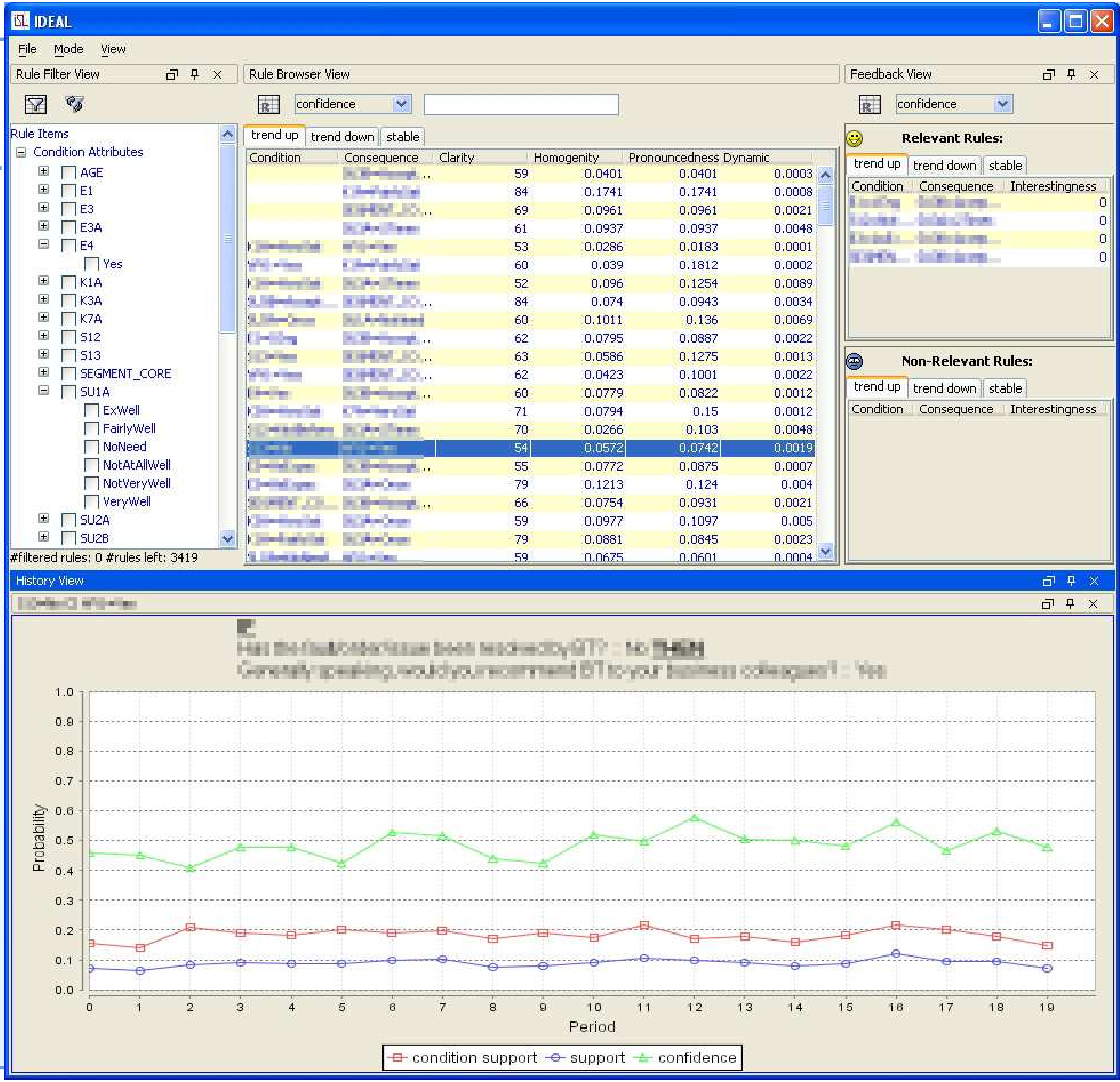
$$dissim_{rs}(\vec{v}, R) = 1 - sim_{rs}(\vec{v}, R)$$

- treatment of non-relevant rules

$$\Psi(\vec{r}, R_{\mathrm{nrel}}) = dissim(\vec{r}, R_{\mathrm{nrel}})$$

- good results for vector similarity with Cosine similarity
- overall: four sets of high-level parameters, which can also be preset

Results


IDEAL
File Mode View
Rule Filter View
Rule Browser View
Feedback View
confidence
Rule Items
Condition Attributes
AGE
E1
E3
E3A
E4
Yes
K1A
K3A
K7A
S12
S13
SEGMENT_CORE
SU1A
ExWell
FairlyWell
NoNeed
NotAtAllWell
NotVeryWell
VeryWell
SU2A
SU2B
#filtered rules: 0 #rules left: 3419
trend up
trend down
stable
Condition
Consequence
Clarity
Homogenity
Pronouncedness
Dynamic
59 0.0401 0.0401 0.0003
84 0.1741 0.1741 0.0008
69 0.0961 0.0961 0.0021
61 0.0937 0.0937 0.0048
53 0.0286 0.0183 0.0001
60 0.039 0.1812 0.0002
52 0.096 0.1254 0.0089
84 0.074 0.0943 0.0034
60 0.1011 0.136 0.0069
62 0.0795 0.0887 0.0022
63 0.0586 0.1275 0.0013
62 0.0423 0.1001 0.0022
60 0.0779 0.0822 0.0012
71 0.0794 0.15 0.0012
70 0.0266 0.103 0.0048
54 0.0572 0.0742 0.0019
55 0.0772 0.0875 0.0007
79 0.1213 0.124 0.004
66 0.0754 0.0931 0.0021
59 0.0977 0.1097 0.005
79 0.0881 0.0845 0.0023
59 0.0675 0.0601 0.0004
Relevant Rules:
Interestingness
0
0
0
0
Non-Relevant Rules:
History View
Probability
Period
condition support
support
confidence

Results - 2

- A relevance feedback system for association rules
- Relevance information is collected from the user.
- High-level weights allow easy adaptation to the user’s search strategy.
- Allows to explore the ruleset intuitively, capturing subjective interestingness.
- Enables on-the-fly change of search strategy.
- No expert user is needed.

Questions?