CUBE BASED SUMMARIES OF LARGE ASSOCIATION RULE SETS

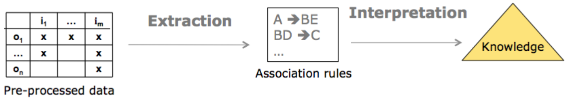
- Marie Ndiaye[1,2], Cheikh Talibouya Diop[2], Arnaud Giacometti[1],
- Patrick Marcel[1], Arnaud Soulet[1]

- [1]LI – Université François Rabelais Tours (France)
- [2]LANI – Université Gaston Berger de Saint-Louis (Sénégal)

This work was partially financed by French Cooperation in Senegal

Context and motivations

- Association rule : implication of the form: $\overset{\text{itemsets}}{\underbrace{\{\text{laptop,bag}\}}_{\text{body}}\rightarrow\underbrace{\{\text{Mouse}\}}_{\text{head}}}$



	i_1	…	i_m
o_1	x	x	x
…	x		x
o_n			x

Pre-processed data

A → BE
BD → C
…

Association rules

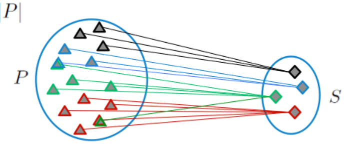
- Data mining algorithms produce large sets of association rules.

Our contributions

- Cube based summaries (CBSs) to explore large sets of association rules
  - The rule sets are summarized according to multiple levels of detail
  - The summaries are represented with cubes
  - OLAP navigational operations can be used to browse the summaries
- An algorithm to generate the most interesting CBS whose size does not exceed a user-specified threshold
  - A quality measure that evaluates the interestingness of CBSs.
  - The obtained summaries initialize the exploration of rule sets

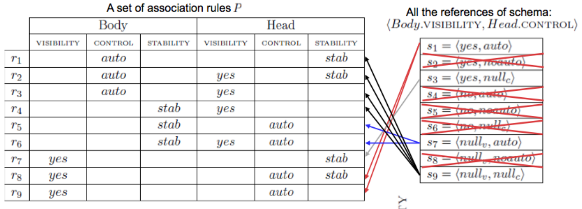
A summary

- Extension of the definition proposed by
  - V. Chandola and V. Kumar: "Summarization - Compressing data into an informative representation" (ICDM'05)
- Two languages of patterns $\mathcal{P}$ and $\mathcal{S}$
- A coverage relation $\lhd$ between $\mathcal{P}$ and $\mathcal{S}$
- A summary of $P \subseteq \mathcal{P}$ is a set of patterns $S \subseteq \mathcal{S}$ such that:
  - Each pattern of P is covered by at least one pattern of S
  - Each pattern of S covers at least one pattern of P
  - $|S| \leq |P|$



Cube based summary (CBS):

- A schema: $\langle Body.\text{VISIBILITY}, Head.\text{CONTROL}\rangle$



A set of association rules P

	Body			Head		
	VISIBILITY	CONTROL	STABILITY	VISIBILITY	CONTROL	STABILITY
r_1		*auto*				*stab*
r_2		*auto*		*yes*		*stab*
r_3		*auto*		*yes*		
r_4			*stab*	*yes*		
r_5			*stab*		*auto*	
r_6			*stab*	*yes*	*auto*	
r_7	*yes*					*stab*
r_8	*yes*				*auto*	*stab*
r_9	*yes*				*auto*	

All the references of schema: $\langle Body.\text{VISIBILITY}, Head.\text{CONTROL}\rangle$

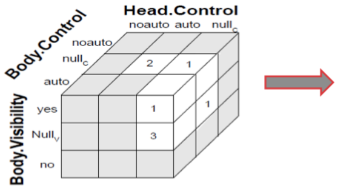
- $s_1 = \langle yes, auto\rangle$
- ~~$s_2 = \langle yes, noauto\rangle$~~
- $s_3 = \langle yes, null_c\rangle$
- ~~$s_4 = \langle no, auto\rangle$~~
- ~~$s_5 = \langle no, noauto\rangle$~~
- ~~$s_6 = \langle no, null_c\rangle$~~
- $s_7 = \langle null_v, auto\rangle$
- ~~$s_8 = \langle null_v, noauto\rangle$~~
- $s_9 = \langle null_v, null_c\rangle$

- CBS S = {s_1,s_3,s_7,s_9}
- It can be represented with cube
- It covers a partition of P and it is a minimal summary
- Several levels of detail

Body.VISIBILITY	*Head*.CONTROL auto	*Head*.CONTROL $null_C$
yes	2	1
$null_V$	2	4

browsing the summaries

- $2^{2^{|A|}}$ summaries for a rule set
- OLAP navigation opérations:
  - Roll-up: deleting an attribute from the schema
  - Drill-down: adding an attribute to the schema
- Example : Roll-up → $Body.\text{CONTROL}$



Body.VISIBILITY	*Head*.CONTROL auto	*Head*.CONTROL $null_C$
yes	2	1
$null_V$	2	4

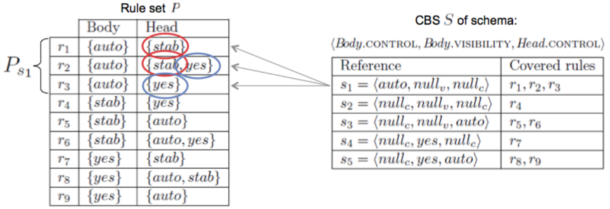
Controling the size of CBS

- Given a schema, we can build a CBS
  - The size of CBSs can indirectly be controlled by choosing their schema: the less there are attributes in the schema, the smaller the CBS is.
- How can we directly control the size of CBSs?
  - Find the most interesting CBS whose size does not exceed a user-specified threshold.
- How can we evaluate the interestingness of CBSs?
  - A quality measure: homogeneity

Homogeneity of a CBS: intuition

- Based on Shannon's conditional entropy
- Evaluates the global homogeneity of the rules covered by the same reference.

$$\phi(P,S) = \sum_{s\in S} \varphi(P_s, s) \leq 0$$



Rule set P

	Body	Head
r_1	$\{auto\}$	$\{stab\}$
r_2	$\{auto\}$	$\{stab, yes\}$
r_3	$\{auto\}$	$\{yes\}$
r_4	$\{stab\}$	$\{yes\}$
r_5	$\{stab\}$	$\{auto\}$
r_6	$\{stab\}$	$\{auto, yes\}$
r_7	$\{yes\}$	$\{stab\}$
r_8	$\{yes\}$	$\{auto, stab\}$
r_9	$\{yes\}$	$\{auto\}$

CBS S of schema:

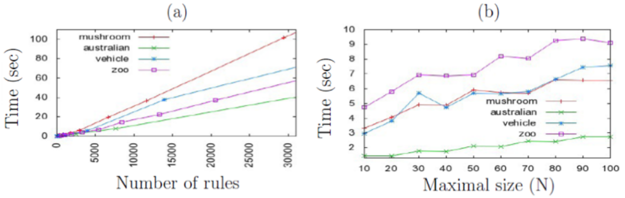
$\langle Body.\text{CONTROL}, Body.\text{VISIBILITY}, Head.\text{CONTROL}\rangle$

Reference	Covered rules
$s_1 = \langle auto, null_v, null_c\rangle$	r_1, r_2, r_3
$s_2 = \langle null_c, null_v, null_c\rangle$	r_4
$s_3 = \langle null_c, null_v, auto\rangle$	r_5, r_6
$s_4 = \langle null_c, yes, null_c\rangle$	r_7
$s_5 = \langle null_c, yes, auto\rangle$	r_8, r_9

- A more specific CBS has a higher homogeneity

Experimental Analysis

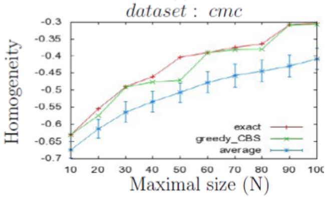
- Runtime Performance of greedy-CBS : 4 datasets



- (a) Increases linearly with the number of rules
- (b) Increases sub-linearly with the maximal size

Experimental Analysis

- Quality of the approximate solutions



- Very close to the optimal solution
- Always over the confidence interval

Conclusion and future works

- A new framework to summarize large sets of association rules.
- A quality measure for CBSs: homogeneity
- An algorithm to generate the most interesting CBS

- Alleviate the constraint of full coverage
- Other experimentations
- Summarize other kind of patterns