Cube Based Summaries of Large Association Rule Sets

Marie Ndiaye[1,2], Cheikh T. Diop[2], Arnaud Giacometti[1],
Patrick Marcel[1], and Arnaud Soulet[1]

[1] Laboratoire d'Informatique, Université François Rabelais Tours,
Antenne Universitaire de Blois, 3 place Jean Jaurès, 41000 Blois (France)
{marie.ndiaye,arnaud.giacometti,
patick.marcel,arnaud.soulet}@univ-tours.fr
[2] Laboratoire d'Analyse Numérique et d'Informatique,
Université Gaston Berger de Saint-Louis, BP 234 Saint-Louis (Senegal)
cheikh-talibouya.diop@ugb.edu.sn

Abstract. A major problem when dealing with association rules post-processing is the huge amount of extracted rules. Several approaches have been implemented to summarize them. However, the obtained summaries are generally difficult to analyse because they suffer from the lack of navigational tools. In this paper, we propose a novel method for summarizing large sets of association rules. Our approach enables to obtain from a rule set, several summaries called Cube Based Summaries (CBSs). We show that the CBSs can be represented as cubes and we give an overview of OLAP [1] navigational operations that can be used to explore them. Moreover, we define a new quality measure called homogeneity, to evaluate the interestingness of CBSs. Finally, we propose an algorithm that generates a relevant CBS w.r.t. a quality measure, to initialize the exploration. The evaluation of our algorithm on benchmarks proves the effectiveness of our approach.

Keywords: Association rules, summary, cubes.

Classical mining algorithms generally produce a huge number of association rules [1] making it difficult to efficiently analyze the discovered rules. Methods that generate generic bases are then proposed to reduce the number of the mined rules [2,3,4]. Unfortunately, this quantity often remains too important. Several methods of summarization and navigation have been proposed to facilitate the exploration of association rule sets. Summarization is a common method for representing huge amounts of patterns [5,6,7,8,9,10]. However, summaries generated by the existing methods are generally difficult to explore. Indeed, they are usually displayed in the form of pattern lists with no further organization. In addition, they suffer from the lack of navigational tools. For the purpose of exploring association rules, these latter are represented with cubes [11,12,13]. In particular, the approach exposed in [13] treats class association rules [14].

[1] On Line Analytical Processing.

L. Cao, J. Zhong, and Y. Feng (Eds.): ADMA 2010, Part I, LNCS 6440, pp. 73–85, 2010.

However, the cubes proposed by those methods don't represent summaries. Indeed, they only depict a portion of the rules which doesn't provide a global and complete view of the whole set of association rules. In this paper, we propose a new approach to explore large sets of association rules.

Our first contribution is the proposal of cube based summaries (CBSs) to summarize large sets of association rules. Theses CBSs can be represented with cubes which enable to depict the rules according to multiple levels of detail and different analytical axes. Our motivation is based on the fact that, by representing rule sets with cubes, we can exploit OLAP navigational operations [15,16] in order to facilitate the analysis of the rules.

Our second contribution is the proposal of a greedy algorithm which generates a relevant CBS not exceeding a user-specified size for a given rule set. Such a CBS can be used to initialize the exploration of the latter. The algorithm is based on a quality measure that evaluates the interestingness of CBSs. We propose a new measure to evaluate the quality of CBSs. It is called homogeneity and is based on the Shannon conditional entropy.

Finally, empirical tests on generic bases show that an interesting CBS can be computed within a reasonable time, even if the size of the initial set of rules is very large. They also show that the quality of a summary generated by our algorithm is close to that of the optimal solution.

The remainder of the paper is organized as follows. Some preliminary definitions and notations are presented in Section 1. In Section 2, we describe CBSs and we give an overview of possible utilizations of OLAP navigational operations. In Section 3, we propose a new quality measure before detailing our summarization algorithm. In Section 4, we present the results of empirical evaluations of our algorithm performed on generic bases of association rules. A state of the art on existing summarization methods is presented in Section 5. We eventually conclude in Section 6 and we expose prospects for future research.

1 Context and Motivation

Summary is practical for representing huge amounts of patterns. It provides a global view of the patterns. It also enables to analyse the extracted patterns in a broader context which highlights relations between them. After recalling the definition of association rule, we propose a summary framework which can be applied not only to association rules, but also to other patterns. Finally, we formulate the problem that we address in the remainder of this paper.

1.1 Association Rules

Let $\mathcal{A}$ be a finite set of attributes such that each attribute $A \in \mathcal{A}$ takes its values in a set $dom(A)$, called the domain of A. In the following, we assume that the domains of the attributes in $\mathcal{A}$ are pairwise disjoint.

Given an attribute $A \in \mathcal{A}$, an *item* defined on A is a value of $dom(A)$.

An *itemset* defined on $\mathcal{A}$ is a set of items $X = \{a_1, ..., a_K\}$ such that $a_k \in dom(A_k)$ for all $k \in \{1, ..., K\}$ and $\{A_1, ..., A_K\} \subseteq \mathcal{A}$. The attribute set $sch(X) = \{A_1, ..., A_K\}$ denotes the schema of X.

Table 1. Association rules

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

An *association rule* defined on $\mathcal{A}$ is a relation $X \Rightarrow Y$ where X and Y are itemsets defined on $\mathcal{A}$ such that $X \cap Y = \emptyset$. X and Y are the body and the head of the rule, respectively. Thereafter, $\mathcal{R}$ denotes the language of all possible association rules.

Let us consider the attribute set $\mathcal{A} = \{\text{CONTROL}, \text{STABILITY}, \text{VISIBILITY}\}$ that describes data about spacecraft landing where $dom(\text{CONTROL}) = \{auto, noauto\}$, $dom(\text{STABILITY}) = \{stab, xstab\}$ and $dom(\text{VISIBILITY}) = \{yes, no\}$. Table 1 shows an excerpt from association rules defined on $\mathcal{A}$. As aforementioned, analysis of rule sets which are often huge in practice requires smaller representations.

1.2 Summary Framework

Even if this paper focuses on summaries of association rules, we present in this section a framework for any language of patterns. Indeed, Definitions 1 and 2 are generalizations of definitions proposed in [6] for itemsets. The notion of summary relies on coverage relation:

Definition 1 (Cover). *Let $(\mathcal{P}, \preceq_{\mathcal{P}})$ and $(\mathcal{S}, \preceq_{\mathcal{S}})$ be two partially ordered pattern languages. A coverage relation over $\mathcal{P} \times \mathcal{S}$, denoted $\lhd$, is a binary relation over $\mathcal{P} \times \mathcal{S}$ such that for all $p \in \mathcal{P}$ and $s \in \mathcal{S}$:*
(i) for all $p' \in \mathcal{P}$, if $p \preceq_{\mathcal{P}} p'$ and $s \lhd p$, then $s \lhd p'$,
(ii) for all $s' \in \mathcal{S}$, if $s' \preceq_{\mathcal{S}} s$ and $s \lhd p$, then $s' \lhd p$.

In our approach, $\preceq_{\mathcal{P}}$ and $\preceq_{\mathcal{S}}$ are specialization relations. The notation $s' \preceq_{\mathcal{S}} s$ means that s' is more general than s and s is more specific than s'. If $\mathcal{P} = \mathcal{S}$ then the specialization becomes a coverage relation. In the following, we consider the specialization relation $\preceq_{\mathcal{R}}$ for association rules described hereafter. Given two association rules $r_1 : X_1 \Rightarrow Y_1$ and $r_2 : X_2 \Rightarrow Y_2$, r_2 is more specific than r_1 ($r_1 \preceq_{\mathcal{R}} r_2$) if $X_1 \subseteq X_2$ and $Y_1 \subseteq Y_2$. For example, $r_2 : \{auto\} \Rightarrow \{stab, yes\}$ is more specific than $r_1 : \{auto\} \Rightarrow \{stab\}$. The specialization relation $\preceq_{\mathcal{R}}$ can also be used as a coverage relation over $\mathcal{R} \times \mathcal{R}$ because the conditions (i) and (ii) of Definition 1 are satisfied. Given a set of patterns $P \subseteq \mathcal{P}$ and a pattern $s \in \mathcal{S}$, $cover(s, P)$ denotes the subset of patterns in P covered by s. Now, let us formalize the notion of summary.

Definition 2 (Summary). *Let $\lhd$ be a coverage relation over $\mathcal{P} \times \mathcal{S}$ and P be a subset of $\mathcal{P}$. A summary of P w.r.t. $\lhd$ is a subset S of $\mathcal{S}$ such that:*
(i) for all pattern p in P, there exists a pattern s of S such that $s \lhd p$,
(ii) each pattern of S covers at least one pattern of P,
(iii) $|S| \leq |P|$.

Subsequently, $\mathcal{S}$ is called a language of summary patterns. Note that most of the existing methods use the same language for P and S [5,6,7,8]. For instance, given the rule set R composed of the rules detailed in Table 1 and the coverage relation $\preceq_{\mathcal{R}}$, the rule set $S = \{r_1 : \{auto\} \Rightarrow \{stab\}, r_{10} : \{\} \Rightarrow \{auto\}, r_{11} : \{\} \Rightarrow \{stab\}, r_{12} : \{\} \Rightarrow \{yes\}\}$ is a summary of R w.r.t. $\preceq_{\mathcal{R}}$. Indeed, conditions (i) and (ii) of Definition 2 are satisfied since $cover(r_1, R) = \{r_1, r_2\}$, $cover(r_{10}, R) = \{r_5, r_6, r_8, r_9\}$, $cover(r_{11}, R) = \{r_1, r_2, r_7, r_8\}$ and $cover(r_{12}, R) = \{r_2, r_3, r_4, r_6\}$. Condition (iii) is also satisfied because S which contains 4 rules is smaller than R with 9 rules. Moreover, if r_1 is removed from S, $S' = \{r_{10}, r_{11}, r_{12}\}$ remains a summary of R because all the rules of R are still covered. But, if one of the other rules is removed, the result is no longer a summary. S' is called a minimal summary of R. This concept is defined below:

Definition 3 (Minimal Summary). *Let $S \subseteq \mathcal{S}$ be a summary of $P \subseteq \mathcal{P}$ w.r.t. $\lhd$. S is minimal if there is no set $S' \subset S$ such that S' is a summary of P.*

1.3 Problem Statement

A pattern set can have several minimal summaries in a given language of summary patterns. For example, the set of rules $\{r_1, r_3, r_4, r_7, r_9\}$ is also a minimal summary of the previous pattern set R. This summary is larger than S given above but it is also relevant because it provides more details about the rules of R. Thus, we would like to navigate between the different summaries of a rule set.

In this paper, we are particularly interested in summaries of association rule sets and we address two problems which are formulated as follows:

1. How should we define a language of summary patterns and a coverage relation that enable to build minimal summaries and to explore effectively large sets of association rules?
2. Which minimal summaries are the most interesting?

Sections 2 and 3 address these problems, respectively.

2 Summarizing Large Sets of Association Rules

To solve the first problem, we propose summaries named cube based summaries (CBSs) that we represent with cubes. Then, we present operations that can be performed on a CBS.

2.1 Cube Based Summary

We introduce in this section CBSs which give smaller representations of large sets of association rules. Fig. 1 depicts a CBS for the set R that contains the association rules presented in Table 1. It is composed of three dimensions which constitute its schema ⟨*Body*.CONTROL, *Body*.VISIBILITY, *Head*.CONTROL⟩. Each dimension corresponds to an attribute. The attributes prefixed with *Head* appear in the head of the rules and those with the prefix *Body* appear in their body. On each dimension, values that belong to the domain of the corresponding attribute

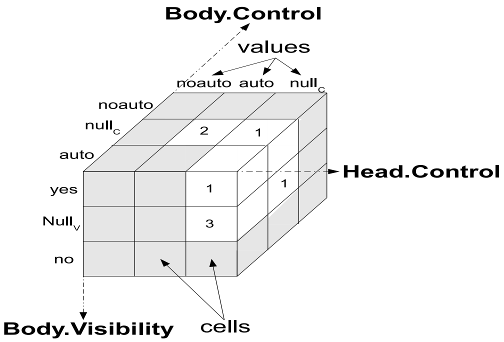

Fig. 1. Representation of a CBS

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

Fig. 2. A roll-up operation

are displayed. The null values $null_C$ and $null_V$ are added on the dimensions to express the absence of value. Each cell of the cube is referenced with a tuple $\langle b_1, b_2, h_1\rangle$ which belongs to $[dom(\text{CONTROL}) \cup \{null_C\}] \times [dom(\text{CONTROL}) \cup \{null_C\}] \times [dom(\text{CONTROL}) \cup \{null_C\}]$. These tuples are called the references defined on the schema $\langle Body.\text{CONTROL}, Body.\text{VISIBILITY}, Head.\text{CONTROL}\rangle$. Thereafter, $\mathcal{S}$ denotes the language of references of all the possible schemas. A reference can cover one or more rules of R and its cell contains the number of rules it covers. The cells filled in grey correspond to the references that cover no rule of R. The values *noauto* of *Body*.CONTROL and *Head*.CONTROL and *no* of *Body*.VISIBILITY can be removed since all the cells associated with them are empty. Our coverage relation is defined over $\mathcal{R} \times \mathcal{S}$. It is based on the specialization relation $\preceq_{\mathcal{R}}$ for association rules (see Section 1) and the specialization relation $\preceq_{\mathcal{S}}$ for references defined hereafter. Given two references $s =< b_1, \ldots, b_I, h_1, \ldots, h_J >$ and $s' =< b'_1, \ldots, b'_K, h'_1, \ldots, h'_L >$, s is more general than s' ($s \preceq_{\mathcal{S}} s'$) if $\{b_1, \ldots, b_I\} \subseteq \{b'_1, \ldots, b'_K\}$ and $\{h_1, \ldots, h_J\} \subseteq \{h'_1, \ldots, h'_L\}$. Considering those specialization relations, we define the following coverage relation that is used in Fig. 1.

Definition 4 (Rule Cover). *Given the language of association rules $\mathcal{R}$ and the language of references $\mathcal{S}$, a reference $s =< b_1, \ldots, b_I, h_1, \ldots, h_J >\in \mathcal{S}$ defined on the schema $\langle Body.B_1, \ldots, Body.B_I, Head.H_1, \ldots, Head.H_J\rangle$ covers a rule $r : X \Rightarrow Y \in \mathcal{R}$ ($s \lhd_{\mathcal{S},\mathcal{R}} r$) if for all $v = b_i$, $i \in \{1, \ldots, I\}$ (resp. $v = h_j$, $j \in \{1, \ldots, J\}$):*

- *when v does not equal to the null value of $Body.B_i$ (resp. $Head.H_j$), X (resp. Y) contains v;*
- *otherwise, X (resp. Y) does not contain an item defined on B_i (resp. H_j).*

For example, the reference $\langle null_C, yes, auto\rangle$ defined on $\langle Body.\text{CONTROL}, Body.\text{VISIBILITY}, Head.\text{CONTROL}\rangle$ covers the rules $r_8 : \{yes\} \Rightarrow \{auto, stab\}$ and $r_9 : \{yes\} \Rightarrow \{auto\}$ because their body and their head contain *yes* and *auto*, respectively. However, $\langle null_C, null_V, auto\rangle$ does not cover $r_9 : \{yes\} \Rightarrow \{auto\}$ because the latter contains in its body the item *yes* which is defined on VISIBILITY.

It can straightforwardly be proved that $\lhd_{\mathcal{S},\mathcal{R}}$ is a coverage relation. Indeed we intuitively observe that every reference that covers a rule $r \in \mathcal{R}$ also covers the specializations of r. And conversely, every rule covered by a reference s is also covered by the generalizations of s. Now, we formalize the notion of CBS.

Definition 5 (Cube Based Summary). *Given a rule set R, the cube based summary of schema $C = \langle Body.B_1, \ldots, Body.B_I, Head.H_1, \ldots, Head.H_J\rangle$ of R, denoted $S_{C,R}$, is the set of all the references defined on C which cover at least one rule of R, w.r.t. $\lhd_{\mathcal{S},\mathcal{R}}$.*

Any CBS $S_{C,R}$ of a rule set R is a summary of R w.r.t. $\lhd_{\mathcal{S},\mathcal{R}}$ because $S_{C,R}$ satisfies the conditions of Definition 2.

The CBS of schema $\langle Body.\text{CONTROL}, Body.\text{VISIBILITY}, Head.\text{CONTROL}\rangle$ depicted in Fig. 1 is composed of the references: $\langle auto, null_V, null_C\rangle$, $\langle null_C, yes, auto\rangle$, $\langle null_C, yes, null_C\rangle$, $\langle null_C, null_V, auto\rangle$ and $\langle null_C, null_V, null_C\rangle$. Note that if at least one of the references detailed above is removed from the CBS, the rule set R is no more covered entirely. The property below generalizes this observation:

Property 1 (Minimality). If $S_{C,R}$ is a CBS of a rule set R, then it is a minimal summary of R w.r.t. $\lhd_{\mathcal{S},\mathcal{R}}$.

Property 1 enables us to ensure that a CBS is not only a summary, but it is also minimal. Therefore, the amount of information presented to the user is reduced as much as possible. In order to respect the requirements regarding the number of pages, the proof of the above property is not exposed.

After proposing the language of references $\mathcal{S}$ and the coverage relation $\lhd_{\mathcal{S},\mathcal{R}}$ that we have used to define CBSs, we show in the next section how CBSs can be explored with OLAP navigational operations.

2.2 Navigational Operations

Classical OLAP navigational operations can be used to explore CBSs, particularly granularity operations, i.e. roll-up and drill-down, which enable to modify the granularity level of the represented data. Roll-up consists in moving from a detailed to a more aggregated level. In our context, it corresponds to the removal of attributes from the schema of a CBS. Drill-down is the reverse of roll-up, it consists in adding attributes to the schema. Fig. 2 shows a result of roll-up performed on the CBS represented in Fig. 1 by removing $Body.\text{CONTROL}$.

Given two CBSs $S_{C_1,R}$ and $S_{C_2,R}$ of a rule set R, if $S_{C_1,R}$ can be obtained by performing a sequence of roll-up from $S_{C_2,R}$, then $S_{C_1,R}$ is more general than $S_{C_2,R}$ and $S_{C_2,R}$ is more specific than $S_{C_1,R}$. The CBS whose schema is empty is the most general and CBS with the schema $\langle Body.B_1, ..., Body.B_I, Head.H_1, ..., Head.H_J\rangle$ where $\{B_1, ..., B_I\} = \{H_1, ..., H_J\} = \mathcal{A}$ is the most specific. Furthermore, the set of the possible CBSs is closed under the granularity operations. Property 2 shows that any CBS can be reached with those operations.

Property 2 (Reachability). Given two distinct CBSs $S_{C_1,R}$ and $S_{C_2,R}$ of a rule set R, there exists a finite sequence of granularity operations $\langle O_1, ..., O_N \rangle$, i.e. roll-up and drill-down operations, such that $S_{C_2,R} = O_1 \circ .... \circ O_N(S_{C_1,R})$.

In conclusion, OLAP navigational operations are well-adapted to explore CBSs. In the next section, we show that some summaries are more suitable than others to perform an effective analysis.

3 Generating an Interesting Cube Based Summary

In this section, we initially focus on the definition of a specific quality measure to evaluate the interestingness of the CBS. Then, we propose an algorithm which finds an approximate solution of the most interesting CBS w.r.t. a quality measure, in order to initialize the exploration of association rules.

3.1 Quality Measure

Given several possible CBSs, the user can hardly identify the most relevant one. Therefore, he needs a measure to evaluate the quality of the CBSs. A quality measure of a summary is a function ϕ that associates with every pair composed of a rule set and a summary, a value in $\mathbb{R}$. In our approach, a CBS is more interesting than its generalizations since it provides more precision about the rules of the set it summarizes. Thus, for assessing effectively the interestingness of CBSs, a quality measure must satisfy the following property:

Property 3 (Monotony). Let ϕ be a quality measure. ϕ is monotone if, for any CBSs $S_{C_1,R}$ and $S_{C_2,R}$ of a ruleset R, if $S_{C_1,R}$ is more specific than $S_{C_2,R}$ then $\phi(R, S_{C_1,R}) \geq \phi(R, S_{C_2,R})$.

Intuitively, if ϕ is monotone then the more specific the CBS is, the higher the measure ϕ. Thus, it is easy to see that the quality of the most general CBS is the smallest. Conversely, the quality of the most specific one is the highest.

Now, we propose a measure called homogeneity that can be used to evaluate the quality of CBSs. Intuitively, the homogeneity reflects the similarity of the rules covered by the same reference. The rules are similar if they have roughly the same items in their head and their body.

Given a ruleset R whose rules are defined on $\mathcal{A}$, let s_j be a reference of a CBS $S_{C,R}$. The similarity between the rules covered by s_j is evaluated by $\sum_{a_i \in \mathcal{I}} p(i,j) \ln[p_j(i)]$ where $\mathcal{I}$ is the set of all the items defined on an attribute of $\mathcal{A}$, $p(i,j) = |\{X \Rightarrow Y \mid (X \Rightarrow Y \in cov(s_j, R)) \wedge (a_i \in X \cup Y)\}|/|R|$ is the joint probability that a rule of R contains a_i and is covered by s_j and $p_j(i) = |\{X \Rightarrow Y \mid (X \Rightarrow Y \in cov(s_j, R)) \wedge (a_i \in X \cup Y)\}|/|cov(s_j, R)|$ is the conditional probability that a rule covered by s_j contains a_i. The homogeneity of $S_{C,R}$ w.r.t. R is given by (1).

$$H(R, S_{C,R}) = \frac{1}{|\mathcal{A}|} \sum_{s_j \in S_{C,R}} \left[\sum_{a_i \in \mathcal{I}} p(i,j) \ln[p_j(i)] \right] \quad (1)$$

$H(R, S_{C,R})$ is a Shannon conditional entropy weighted with the size of $\mathcal{A}$. More details on conditional entropy properties can be found in [17]. Furthermore, the homogeneity satisfies Property 3. Its value is negative or null and it corresponds to 0 for the most specific CBS.

For example, let us consider the item $a_i = yes$ and the reference $s_j = \langle null_C, yes, auto\rangle$ of the CBS $S_{C_1,R}$ that summarizes the rule set of Fig. 1 where $C_1 = \langle \mathit{Body}.\textsc{control}, \mathit{Body}.\textsc{visibility}, \mathit{Head}.\textsc{control}\rangle$. Knowing that $cov(s_j, R) = \{r_8, r_9\}$ with $r_8 : \{yes\} \Rightarrow \{auto, stab\}$ and $r_9 : \{yes\} \Rightarrow \{auto\}$, we have $p(i,j) = \frac{|\{r_8,r_9\}|}{|R|} = \frac{2}{9}$ and $p_j(i) = \frac{|\{r_8,r_9\}|}{|\{r_8,r_9\}|} = \frac{2}{2}$. When we measure the homogeneity of the CBSs $S_{C_1,R}$ and $S_{C_2,R}$ represented in Fig. 1 and Fig. 2, respectively, we obtain $H(R, S_{C_1,R}) = -0,24$ and $H(R, S_{C_2,R}) = -0,3$. We observe that Property 3 is satisfied since $S_{C_1,R}$ is more specific than $S_{C_2,R}$ and $H(R, S_{C_1,R}) > H(R, S_{C_2,R})$.

3.2 Algorithm for Finding an Interesting Cube Based Summary

This section aims at initializing navigation by automatically searching in the space of the possible CBSs, an interesting one from which the user can begin the exploration. The problem can be formulated as follows: given a set of rules R extracted from a dataset of schema $\mathcal{A}$ and a quality measure ϕ, find the most interesting CBS $S_{C^*,R}$ (w.r.t. ϕ) whose size is smaller than a fixed threshold N. This problem is NP-complete. Its NP-completeness can be shown by using the knapsack problem. Hence, we propose an approximate solution computed with a greedy algorithm. We begin from the most general CBS $S_{C_0,R}$ where $C_0 = \langle\rangle$ and we use granularity navigational operations to explore the space of the possible CBSs. Recall that this space is closed under the granularity operations. More precisely, we use the two following drill-down operations:

Algorithm 1. greedy_CBS

```
Input: $R$ {Set of rules}, $\mathcal{A}$ {The set of attributes}
   and $N$ {The maximal size of the summary}
Output: {A cube based summary}
$C_0 = \langle\rangle$, $i = 0$ {Initializations}
repeat
   $i = i + 1$
   $C_i = C_{i-1}$
   for all $\mathcal{O} \in \{AddToHead, AddToBody\}$ do
      for all $A \in \mathcal{A}$ do
         $C = \mathcal{O}(S_{C_{i-1},R}, A)$
         if $\phi(R, S_{C,R}) > \phi(R, S_{C_i,R})$ and $|C_i| \leq N$ then
            $C_i = C$
         end if
      end for
   end for
until $\phi(R, S_{C_{i-1},R}) = \phi(R, S_{C_i,R})$
return $S_{C_i,R}$
```

AddToHead and *AddToBody*. Given an attribute A, $AddToHead(S_{C,R}, A)$ and $AddToBody(S_{C,R}, A)$ consist in adding $Head.A$ and $Body.A$ in C, respectively. The CBSs whose schema result from either *AddToHead* or *AddToBody* operations are more specific than $S_{C,R}$. Therefore, their quality is better than H(R, $S_{C,R}$). In the worst case, they have the same quality as $S_{C,R}$.

Algorithm 1 seeks an approached solution of $S_{C^*,R}$. It consists in choosing at each step i an attribute from $\mathcal{A}$ which will be added to $S_{C_{i-1},R}$ to obtain $S_{C_i,R}$. The attribute retained at step i is selected such that the CBS $S_{C_i,R}$ has the greatest quality compared to the summaries obtained by adding one of the other attributes of $\mathcal{A}$ (line 7 to 9). The addition of attributes stops when the quality of the CBS obtained in the previous step cannot be improved.

4 Experimental Analysis

In this section, we study the performance of our proposed approach on benchmarks. We use discretized datasets[2] resulting from the UCI Machine Learning repository[3]. We perform our experimentations on generic bases of association rules extracted with the CHARM[4] [18] algorithm. Our summarization algorithm is implemented in Java. All the experiments are executed on a computer Intel dual core 2GHz with 2GB-memory and running Windows Vista. We evaluate the summarization performance in terms of computation time and homogeneity.

4.1 Runtime Performance

We conducted our first experiment using the data sets shown in the table in Fig. 3. For each data set, we generate several rule sets by varying the support level (*minsup*) where the confidence level (*minconf*) is fixed to 50%.

Fig. 3a reports the computation time of the CBSs generated with the maximal size $N = 50$ w.r.t. the number of rules contained in the generic bases. We first notice that the runtime does not exceed 120 seconds even if the rule set contains more than $30,000$ rules. Moreover, the slope of the curves depends on the number of attributes in $\mathcal{A}$. The larger the number of attributes, the higher the slope of the curve. For instance, the curve of *mushroom* has the greatest slope because this dataset contains more attributes (23 attributes) than the others.

Fig. 3b plots the computation time of CBSs when their maximal size N varies between 10 and 100. For this experiment, the rule sets are generated with $minconf = 50\%$ and $minsup = 25\%$ for *mushroom* and *zoo* and $minconf = 50\%$ and $minsup = 15\%$ for *australian* and *vehicle*. We observe that for all the datasets, the computation time sublinearly increases with N. Indeed, in the algorithm implementation, each attribute added by using an operation is not tested again with this operation.

[2] `users.info.unicaen.fr/~frioult/uci`

[3] `mlearn.ics.uci.edu/MLRepository.html`

[4] `www.cs.rpi.edu/~zaki/software/`

			minsup										
		$\lvert\mathcal{A}\rvert$	5%	8%	10%	15%	20%	25%	30%	35%	40%	45%	50%
Dataset	mushroom	23	-	38406	29441	11629	6681	2915	1732	838	390	227	110
	vehicle	19		31873	13890	3899	1066	339	52	4	0	0	0
	australian	15	39060	-	7573	2437	1019	486	247	124	62	23	9
	zoo	17	31053	-	20583	13253	8446	5382	3283	1864	957	569	300

Number of rules generated from datasets according to *minsup* where $minconf = 50\%$

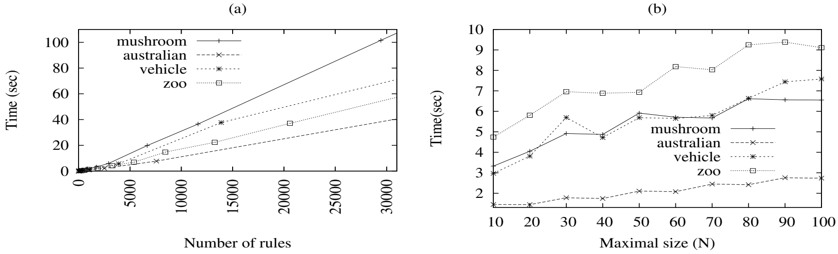

Fig. 3. Execution time

4.2 Quality of the Approximate Solution

This section aims at evaluating the quality of our approximate solution. Thereby, given a rule set R and a maximal size N, we compare the homogeneity of CBSs resulting from 3 approaches:

- **Greedy solution (`greedy_CBS`):** Our algorithm produces an approximate solution.
- **Optimal solution (`exact`):** A naive algorithm enumerates all the CBSs in order to return the optimal solution $S_{C^*,R}$.
- **Average solution (`average`):** A naive algorithm enumerates all the summaries in order to compute the average homogeneity (and the standard deviation) of the most specific summaries (not exceeding N references).

Of course, the algorithm that computes the optimal solution is very costly and fails on datasets containing a large number of attributes. For this purpose, we perform experiments on the sets of association rules mined stemming from small datasets: *cmc*, *glass* and *tic-tac-toe*. The table in Fig. 4 details the number of rules for each generic base according to the dataset, the minimal support and confidence thresholds.

Fig. 4 plots the homogeneity of the three approaches detailed above (i.e. `greedy_CBS`, `exact` and `average`) for each generic base w.r.t. the maximal size of the CBSs. Let us note that we also report the confidence interval of the `average` curve. As expected, we observe that homogeneity logarithmically increases with the maximal size N for the three approaches. This is because the more N increases, the more the generated CBS is specific. Furthermore, we observe that the homogeneity of our summaries is close to that of the optimal summaries for all the experiments. Even if the threshold N increases, the distance between the approximate solution and the optimal one remains moderate. In Fig. 4a,

	Dataset		
	cmc	glass	tic-tac-toe
$\|\mathcal{A}\|$	10	10	10
minsup	8%	6%	5%
minconf	80%	50%	50%
Number of rules	1116	1210	1299

Data description

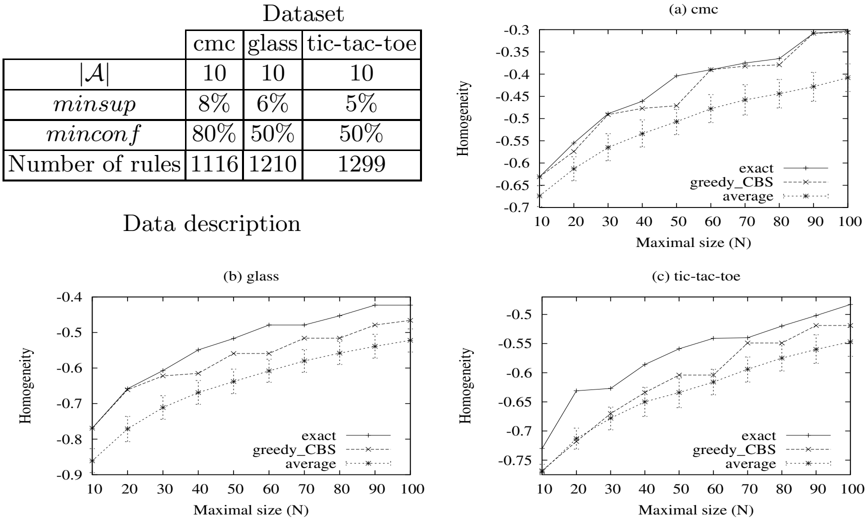

Fig. 4. Homogeneity of CBSs

we notice that for several values of N (e.g., 60 or 90), the homogeneity of our summaries is even optimal. We also notice that the homogeneity of summaries generated by `greedy_CBS` is always greater than that of the summaries generated by `average`. More interestingly, our algorithm finds really relevant CBSs (w.r.t. the homogeneity) because the homogeneity of `greedy_CBS` is often above the confidence interval.

5 Related Work

In the past few years, summarization of pattern sets has been addressed following several approaches. Table 2 highlights some characteristics of the approaches closest to ours [10,6,5,8,7,9] . We comment the table below.

Language of patterns and language of summary patterns: The considered patterns in practically all the cited papers are itemsets, except in [9] where the authors address the problem of summarizing association rules. Furthermore, the patterns of the summaries generally belong to the same language as those of the initial set. However, in [10] authors use pattern profiles which describe the itemsets and approximate their support. In [9], the patterns of the summaries are class association rules [14], i.e. association rules whose head are restricted to a class item.

Coverage: The methods proposed in [10,6,7,9] produce summaries that cover the entire set, i.e. each pattern is covered by at least a pattern of the summary, contrary to other methods which provide summaries that approximately cover the set of association rules [5,8]. The Summaries of the former methods are summaries in the sense of definition 2.

Table 2. Methods for summarizing sets of patterns

Reference	[9]	[6]	[5]	[10]	[7]	[8]	Our method
$\mathcal{P}$	rules	itemsets	itemsets	itemsets	itemsets	itemsets	rules
$\mathcal{S}$	class rules	itemsets	itemsets	profiles	itemsets	itemsets	references
Regeneration	no	no	patterns	frequency	frequency	frequency	no
Coverage	entirely	entirely	partially	entirely	entirely	partially	entirely
Measure ($\star$)	no	CG, IL	NPC	RE	RE	IL	H
Representation	no	no	no	no	no	no	yes
Navigation	no	no	no	no	no	no	yes

$\mathcal{P}$: Language of patterns $\mathcal{S}$: Language of summary patterns
($\star$) CG: compaction gain, RE: restoration error, IL: information loss, H: homogeneity
NPC: number of patterns covered

Regeneration: The main objective of some approaches is to build summaries so that they can be used to regenerate the original itemsets [5] or their frequency [8,7,10] whereas the others are interested in the representation of the patterns for future exploration [6,9]. Our approach is in the latter category.

Measure: Measures of restoration error [10,7] and information loss [8] are used to evaluate the error generated when the frequencies of the patterns can be approximated from a summary. The measure used in [5] calculates the size of the subset of patterns which are really covered by at least one pattern of the summary. In [6], the authors apply a measure called compaction gain to evaluate the compression ratio of a pattern set w.r.t. its summary. They also use an information loss measure to assess the quantity of information contained in the patterns and absent from the patterns of the summary which cover them. The measures which evaluate the regeneration error and the quantity of covered patterns are not adapted to our approach. Indeed, the CBSs are not intended to regenerate information and they always cover the entire pattern set.

Representation and navigation: Our approach enables to explore summaries represented with cubes by using OLAP operations. Contrary to us, the approaches cited above don't propose a representation for the summaries they build. Furthermore, those approaches suffer from the lack of exploration techniques.

6 Conclusion

We have proposed in this paper a new framework to summarize large sets of association rules. Our summaries, called Cube Based Summaries (CBSs) can be represented with cubes and explored using OLAP navigational operations. An algorithm is presented to generate an interesting summary w.r.t. a quality measure of summary in order to initialize the rules exploration. Finally, our experiments prove the feasibility of our approach.

We project to propose a generalization of our approach by using our summary framework, in order to summarize other types of patterns. Furthermore, the space of CBSs is not closed under the OLAP selection operations because they produce portions of summaries. We aim to define a selection operation that provides CBSs like granularity operations.

References

1. Agrawal, R., Imieliński, T., Swami, A.: Mining association rules between sets of items in large databases. In: SIGMOD, pp. 207–216. ACM, New York (1993)
2. Liu, B., Hsu, W., Ma, Y.: Pruning and summarizing the discovered associations. In: KDD 1999, pp. 125–134 (1999)
3. Srikant, R., Vu, Q., Agrawal, R.: Mining association rules with item constraints. In: KDD 1997, pp. 67–73 (1997)
4. Zaki, M.J.: Generating non-redundant association rules. In: KDD 2000, pp. 34–43 (2000)
5. Afrati, F., Gionis, A., Mannila, H.: Approximating a collection of frequent sets. In: KDD 2004, pp. 12–19 (2004)
6. Chandola, V., Kumar, V.: Summarization — compressing data into an informative representation. In: ICDM 2005, pp. 98–105 (2005)
7. Jin, R., Abu-Ata, M., Xiang, Y., Ruan, N.: Effective and efficient itemset pattern summarization: regression-based approaches. In: KDD 2008, pp. 399–407 (2008)
8. Mielikäinen, T., Mannila, H.: The pattern ordering problem. In: Lavrač, N., Gamberger, D., Todorovski, L., Blockeel, H. (eds.) PKDD 2003. LNCS (LNAI), vol. 2838, pp. 327–338. Springer, Heidelberg (2003)
9. Ordonez, C., Ezquerra, N., Santana, C.: Constraining and summarizing association rules in medical data. Knowledge and Information Systems 9, 259–283 (2006)
10. Yan, X., Cheng, H., Han, J., Xin, D.: Summarizing itemset patterns: a profile-based approach. In: KDD 2005, pp. 314–323 (2005)
11. Blumenstock, A., Schweiggert, F., Müller, M., Lanquillon, C.: Rule cubes for causal investigations. Knowledge and Information Systems 18, 109–132 (2009)
12. Boulicaut, J.F., Marcel, P., Rigotti, C.: Query driven knowledge discovery in multidimensional data. In: DOLAP 1999, pp. 87–93 (1999)
13. Liu, B., Zhao, K., Benkler, J., Xiao, W.: Rule interestingness analysis using olap operations. In: KDD 2006, pp. 297–306 (2006)
14. Hu, K., Lu, Y., Zhou, L., Shi, C.: Integrating classification and association rule mining: A concept lattice framework. In: Zhong, N., Skowron, A., Ohsuga, S. (eds.) RSFDGrC 1999. LNCS (LNAI), vol. 1711, pp. 443–447. Springer, Heidelberg (1999)
15. Chaudhuri, S., Dayal, U.: An overview of data warehousing and olap technology. SIGMOD Rec. 26(1), 65–74 (1997)
16. Romero, O., Abelló, A.: On the need of a reference algebra for olap. In: Song, I.-Y., Eder, J., Nguyen, T.M. (eds.) DaWaK 2007. LNCS, vol. 4654, pp. 99–110. Springer, Heidelberg (2007)
17. Shannon, C.E.: A mathematical theory of communication. SIGMOBILE Mob. Comput. Commun. Rev. 5(1), 3–55 (2001)
18. Zaki, M.J., Hsiao, C.J.: Charm: An efficient algorithm for closed itemset mining. In: SDM, pp. 457–473 (2002)