Database summarization approach based on description logic theory

Amel TRIKI, Yann POLLET, Mohamed BEN AHMED
RIADI-GDL Laboratory, CEDRIC-CNAM Laboratory, RIADI-GDL Laboratory
amel.triki@riadi.rnu.tn, pollet@cnam.fr, mohamed.benahmed@riadi.rnu.tn

Abstract- **In this paper, we propose a new approach of database summarization. Our proposal consists in building a set of summaries that gives many levels of granularity.**
The main contribution of our work consists in giving a generic approach, based on description logic language, which operates on both the schema and the database content.
The summarization process leads to building a lattice of summaries where each one gives a certain measure of precision.
Our proposal offers a generic setting in which current summarization techniques can be considered as particular cases.
KEYWORDS: database summarization, description logic, summaries lattice, granularity.

I. INTRODUCTION

The increasing evolution of databases volumes raises many difficulties to mine and hold large amounts of data.

Therefore, it is interesting to have an approach that gives a concise and intelligible representation of data through the development of a database summary in order to facilitate their analysis.

In this objective, our paper aims at proposing a new approach of database summarization based on description logic theory.
Our proposal consists in building a set of summaries that gives many levels of granularity. An overview of the method is given in the paper.

The remainder of this paper is structured as follows; in the second section, we give a general overview of the main related work. The third section exposes our basic assumptions, our definition of database summarization and the main issues related to this topic. In the fourth section, we detail our approach of database summarization. In the fifth section, we present the main database summaries properties. Finally, we present our conclusion and future work.

II. RELATED WORK

Various approaches has been investigated to reduce databases volumes which can be categorised mainly in four different ways: methods based on unary operators (vertical and horizontal reduction), approaches related to multidimensional databases such OLAP and QuotientCube, methods based on statistic and symbolic techniques and appoaches based on fuzzy set theory.

The first summarization way is mainly based on Projection and Selection techniques.

Projection is a vertical reduction which removes some attributes whereas selection: is a horizontal reduction by removing some tuples from the database. [1]
These two techniques are relatively accomodating as they reduce considerably the database volume, but they present two major disadvantages. Firstly, the data amount degrades rapidly so that it is not possible to have graduated information in the obtained summary. Secondly, deduced information like the one found by the group by operator can not be acquired.

This last point of view can be found in OLAP (On Line Analytical Processing) and multidimensional databases.
These methods have drawn special interest since they allow capturing and presenting data as arrays that can be arranged in multiple dimensions.

Quotient Cube, a sort of Group by generalization, aims also at summarizing important volumes. [2]

The third summarization approch consists in "statistic methods" which consists globally in replacing a set of tuples by a statistic indicator. This method presents an unshakable asset because volumes are reduced considerably; however, it is not adapted to nominal data and doesn't take into account data semantics.
This explains the emergency of techniques based on symbolic objects. The goal of such methods is to generate a higher level database containing "macro objects" standing for a set of individual objects having a certain level of similarity in their description. We point out here on the researches of E. Diday in this field [3]. These approaches work mainly on a set of tuples, not on the whole database.

The last category of summarization approaches uses fuzzy set techniques. In this context, we can mention mainly the framework SaintEtiq of on line database summarization. [4] [5] [6] [7] [8] [9]
We can also mention the approach aiming at building a schema summarization [10] [11] [12].

Most of the existing approches use fuzzy set techniques but did not all lead to significant and measurables results. [13][14][15][16]
Our work contribution is to give a different point of view and different approach of summarization based on description logic theory in which these mentioned techniques can be considered as particular cases of summary.
We detail in the next section our database summarization approach.

III. DATABASE SUMMARIZATION

A. Basic assumptions

In all the paper, we assume that the database is given by an UML class diagram composed of classes and relations.

A database B can be expressed as a triplet <I, R, E> where **I** designates the database **intension** which can be represented by a set of classes {C1, C2, …, Ci…, Cn}, n, i ∈ **N** (natural numbers), Ci ∈ $\mathcal{C}$ (all the possible classes)

Each class is defined by a set of attributes {A1, A2,…, Ai,…, An} n, i ∈ **N** and Ai ∈ $\mathcal{A}$ (class attributes) where each attribute has a domain *DA*..

We notice that DA designates all the values that can be taken by the attributes which can be either quantitative or qualitative.

R designates the relations {R1, R2,…, Ri, …, Rm} m, i ∈ **N** and R ∈ $\mathcal{R}$ (all the possible relations : association, generalization,…).

The intension and relations expresses the database schema.

E designates the **extension** which represents all the class instances (the database tuples)

B. Defining database summary

A database summary can be defined as a concise representation of a set of structured data [17].

The database summarization can concern three aspects which are:

- The database schema that concerns the classes and the relations between them,
- The database attributes,
- The database tuples.

We emphasise that summarizing a database doesn't mean necessarily reducing the number of classes or attributes, it is possible that the summarization process implies the creation of new classes or attributes that will include concised information.

C. Summarization issue

Database summarization raises many issues which can be described as follows:

- The approach must be generic in the sense that it has to be independent from the database model (relational, object oriented,...), this explains our assumption of an UML class diagram as an initial database structure,
- The approach must take into account the database evolution by adding, removing and modifying either the schema or the database attributes and tuples. However, this point is, at this stage, beyond the scope of this paper,
- The summarization approach must operate as well on the structure that on the content.

Taking account of all these reflections, we introduce in the next section an overview of our database summarization approach.

IV. OVERVIEW OF THE DATABASE SUMMARIZATION APPROACH

One of the main problems encountred is to find a generic proposal that deals with both the database schema and instances.

We have to notice that the database summarization can concern the schema, the class attributes and the instances. So, we aim at finding a generic formalism that merge both the entension and the intension.

D. Database expression in description logic

Description logic (DL) is a knowledge description represetation language. DL distinguishes between the terminological (TBOX) and the assertional description (ABOX). TBOX contains a description of the concepts hierarchies and the relations between them, whereas the ABOX details where individuals belongs to in the hierarchy [14].

In the following subsections, we detail the TBOX and ABOX related to our case study.

D.1 TBOX

The TBOX contains the definition of the different classes and the possible roles (relations) between them according to the following model:

Concepts :

Class <= T (Top concept)

MotherClass = Class $\cap \exists$ generalise.Class

Roles :

generalize, associate, aggregate, …

The list is not exhaustive, we only give samples to explain the way we transform our database schema to logic expressions.

D.2. ABOX

Once all the concepts and roles specified, we define the assertions as the example given in figure 1:

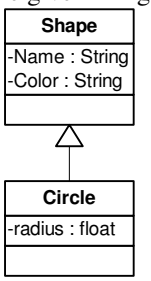

Fig. 1. Subset of a database expressed in UML

Class (Circle)

Class (Shape)

generalize (Circle, Shape)

So with the defintion of the TBOX, it will be deduced that Shape is a MotherClass.

We draw attention to the fact that to express effectively the database in description logic, we opt for using the Web Ontology Language OWL DL since it seems the most suitable representation language. So, an excerpt of the expression of the example described above in OWL DL is given as follows:

```
<owl: Class rdf: ID =”Shape”/>
<owl: Class rdf : ID =”Circle”>
   <owl: subClassOf rdf: resource= “#Shape”/>
</owl: Class>
<owl : DatatypeProperty rdf : ID = “radius”>
  <owl : domain rdf : resource = “#Circle”/>
  <owl : range rdf : resource = “&xsd; float”/>
</owl : DatatypeProperty>
```

Once the database schema and instances are specified in description logic, some rules are defined to transform the database in order to summarize it. This aspect is detailed in the next subsection.

E. Database transformations

As explained in the last subsection, the database is formalised in description logic language so that the schema and instances are expressed by logic expressions.

A summary is therefore a new set of logical expressions which can be deduced from the original database.

Database transformation includes many aspects; the database restructuring, attributes reducing or grouping and instances reduction.

In fact, some implications have to be defined as subsumption which denotes a hierarchical relation, so the former classes (defined in the database schema) would be replaced by generalized classes but formalized in a description logic language.

So, for instance, in the example given earlier, the class Circle would be replaced by the class Shape.

Concerning attributes, two possibilities can be considered.

With regard to the first possibility, each attribute will have a certain weighting (assigned by the future database user) that indicates its relevance as attributes are context dependent. So, some attributes can therefore be omitted.

The second point of view is that we use clustering algorithm to group attributes having certain proximity. In this case, we have also to define a parameter related to attributes weightings.

Finally, concerning instances, we aim at using symbolic data techniques to describe set of tuples by complex and multivaluated data.

For instance, summarizing the values of a numeric type attribute means the definition of a symbolic object of type interval which boundaries are the smallest and the biggest values that can be taken by the attribute.

In the case of a nominal type attribute, the summarization process implies the generation of a symbolic object which is a set of all the attribute nominal values.

To illustrate this point, we consider the following sample of the database (Table I) which contains the class Circle tuples:

TABLE I
EXCERPT OF THE CLASS CIRCLE TUPLES

Name	Circle1	Circle2	Circle3	Circle4	Circle5	Circle6
Color	Green	Green	Red	Green	Red	Red
Radius	1	0.5	1	0.7	0.5	0.3

A first way of summarization is to group the different tuples per color, so, the radius value is replaced by a symbolic object which is an interval where the boundaries are the minimum and the maximum values of the radius as illustrates in table II.

TABLE II
SUMMARY OF THE TUPLES GROUP BY COLOR

Color	Green	Red
Radius	[0.5, 1]	[0.3, 1]

The same tuples can also be summarized differently according to the radius value as shown in the third table.

TABLE III
SUMMARY OF THE TUPLES GROUP BY RADIUS

Radius	0.3	0.5	0.7	1
Color	{Red}	{Green, Red}	{Green}	{Green, Red}

We can also have other combinations; it depends on the end user will.

So, by applying the transformations introduced above, a database can generate various summarized databases with different points of view and levels of granularity. In the next subsection, we prove that the obtained summaries can be organised within a lattice structure.

F. Building summaries lattice

For the rest of the paper, we designate the obtained summaries by S (all the possible summaries) and $\leq$ the subsumption relation that interrelates the different summaries.

We consider P(*S*) as the parts of the summaries: P(S) = {Si/Si $\in$ *S*} and S1, S2, S3 three summaries included in *S*.

It is possible to notice that:

- $\leq$ is reflexive: $\forall$ S1$\in$ P(*S*) : S1 $\leq$S1
- $\leq$ is antisymmetric: $\forall$ S1$\in$ P(*S*) , $\forall$ S2$\in$ P(*S*) : S1 $\leq$ S2 and S2 $\leq$ S1 $\Rightarrow$ S1 = S2
- $\leq$ is transitive : $\forall$ S1, S2, S3$\in$ P(*S*)^3 : S1 $\leq$ S2 and S2 $\leq$ S3 $\Rightarrow$ S1 $\leq$ S3

So, subsumption ($\leq$) defines a partially ordered set.

(P(*S*), $\leq$) defines a lattice. In fact, it has a supremum and an infimum, where the supremum is the initial database and the infimum is the empty set.

V. DATABASE SUMMARY PROPERTIES

We demonstrated earlier that the database summaries would be organised within a lattice structure, so, one of the main points to deal with is how to characterize the different summaries.

At first, we can distinguish mainly two criteria: *informativity* and *consistency* which can be considered as inversely proportional.

The more the summary is informative, the less it is consistent.

So, informativity measures how information can be given by a summary, as far as we browse the lattice to the infimum, as far as the information is degraded.

Consistency measures data volume necessary to define the database granularity.

Informativity combined with consistency can be considered as a criterion to choose a certain summary precision.

VI. CONCLUSION AND FUTURE WORK

In this paper we proposed a description logic approach to build database summaries lattice.

Our proposal method consists globally in three steps.

The first step consists in formalizing the database using description logic (DL) language.

This choice has been motivated by the ability of DL in expressing both the intension and the extension of the database in the same formalism.

The second step consists in defining rules that give a simplified view of the original database. These rules give the possibility of reducing the schema, attributes and the instances of the database.

So, in other words, this step consists in applying consecutive operators that will progressively reduce the data volume. But, it is important to underline that such technique doesn't allow to have a summary like the one obtained by applying the "group by" operator for instance. This remark suggests us to complete our proposal by investigating other solutions.

We mean that we have to build a method that doesn't simplify the database in only one stage. It is possible to add new classes or attributes in our database, as a first step, to obtain aggregated information relevant for the summary, then, in a second step, we reduce the database schema as explained earlier.

The third step of our method consists in organizing the different summaries obtained after the transformations in a lattice structure.

So, browsing the summaries lattice gives the opportunity to choose a summary having a certain granularity according to a precise criterion.

As future work, we aim at detailing more all the steps of our approach and deeping the simplification method as mentioned earlier.

REFERENCES

[1] J.F. Roddick and al. "Methods and Interpretation of Database Summarisation". DEXA. Pp. 604-615.1999.

[2] J.Gray and al."Data Cube: A Relational Aggregation Operator Generalizing Group-By, Cross-Tab, and Sub-Total". Computing Research Repository (CoRR). 1999.

[3] E. DIDAY. " De la statistique des données à la statistique des connaissances : avancées récentes en Analyse des Données Symboliques". EGC. p 703. 2005

[4] G. Raschia, N. Mouadib. "SEQ: a fuzzy set-based approach to database summarization. Fuzzy Sets and Systems". Vo 129(2): pp. 137-162 2002.

[5] G. Raschia. "Linguistic Summarization of a relation with Fuzzy Background Knowledge". *BDA*. 2001

[6] G. Raschia, N. Mouadib. "Using Fuzzy Labels As Background Knowledge for Linguistic Summarization of Databases". *FUZZ-IEEE*. Pp.1372-1375. 2001.

[7] R. Saint Paul, G. Raschia, N. Mouadib. "Database Summarization: The SaintEtiQ System". *ICDE.* Pp. 1475-1476. 2007.

[8] R. Saint Paul, G. Raschia, N. Mouadib. "Résumé généraliste de bases de données ". *BDA*. 2005

[9] R. Saint Paul, G. Raschia, N. Mouadib. . "General Purpose Database Summarization". VLDB. Pp. 733-744. 2005.

[10] D. Lee, M. Kim. "Database summarization using fuzzy ISA hierarchies". *IEEE Transactions on Systems*, Man, and Cybernetics, Part B 27(1):. Pp. 68-78. 1997.

[11] D. Lee, M. Kim. "Discovering Database Summaries through Refinements of Fuzzy Hypotheses." *ICDE 1994*. Pp.223-230. 1994.

[12] D. Lee, M. Kim. "A Hypothesis Refinement Method for Summary Discovery in Databases". *CIKM*.. pp. 274-282. 1993.

[13] D. Dubois, H. Prade. "Fuzzy sets in data summaries—outline of a new approach". In *Proceedings 8th Int. Conf. on Information Processing and Managament of Uncertainty in Knowledge-based.* (2000).

[14] K. Janusz. "Fuzzy logic for linguistic summarization of databases". *In Proceedings of the 8th International Conference on Fuzzy Systems (FUZZ-IEEE'99), Vol. 1 (pp. 813–818). Systems (IPMU'2000*), Vol. 2 (pp. 1035–1040). 2000.

[15] L. Naoum. "Un modèle multidimensionnel pour un processus d'analyse en ligne de résumés flous", Ph. D. Thesis, Université de Nantes, October 2006.

[16] L. Naoum. « Représentation de résumés de base de données par prototypes flous ». *In: 14es Rencontres Francophones sur la Logique Floue et ses Applications* (LFA),.2006

[17] A. Napoli. *Une introduction aux logiques de description*. Technical Report, n°3314, INRIA, 1997.