

QUERYING FUZZY SUMMARIES OF DATABASES: UNARY OPERATORS AND THEIR PROPERTIES

Lamiaa Naoum Guillaume Raschia Noureddine Mouaddib
Lamiaa.Naoum@univ-nantes.fr Guillaume.Raschia@univ-nantes.fr Noureddine.Mouaddib@univ-nantes.fr

LINA

Université de Nantes
2, rue de la Houssinière,
BP 92208, 44322
Nantes Cedex 3 - FRANCE

Abstract

In this paper we propose a general framework to explore and analyse database summaries built from massive data sets. Summaries are multidimensional vague concepts, represented by a set of fuzzy terms provided on each attribute as domain knowledge. We define a logical data model called *summary partition*, as cubes do in OLAP systems. We then propose a set of unary operators over this data model. We also study their properties.

Keywords: database summarization, vague concept, OLAP cubes.

1 Introduction

Because of the ever increasing amount of information stored each day into databases, users can no longer have an exploratory approach for visualizing, querying and analyzing their data without facing the problem often referred as 'Information Overload'. Hence, as mentioned in [5], the data summarization paradigm has become "a ubiquitous requirement for a variety of application environments, including corporate data warehouses, network-traffic monitoring and large socio-economic or demographic surveys".

On-Line Analytical Processes (OLAP) and multidimensional databases are arising great interest from the summarization task point of view, since they allow an end-user to query, visualize and access part of the database using cubes of aggregate values computed from raw data. Therefore, in complement to these well-known OLAP approaches, a fuzzy set-based summariza-

tion method called SaintEtiQ have been proposed by [11] providing summaries which cover parts of the primary database. Since interpretation and exploration of summaries is a main goal of summarization, the symbolic/numerical interface provided by tools of the Zadeh's fuzzy set theory [14], are the fundamental background of all the approaches to linguistic summarization. Significant works have been done in this area, for instance by Yager [13], Rasmussen and Yager [12], Kacprzyk [7], Bosc et al. [3], Cubero et al. [4], Dubois and Prade [10]. Other interesting works are done in fuzzy multidimensional model by Laurent [9], Delgado et al. [6] and Blanco et al. [2]. Inspired by research works in [1] where the authors define a set of algebraic operators for OLAP cubes. We propose the adaptation of this algebra to linguistic fuzzy summaries which provide to the user some intentional descriptions of part of the data set. To do that we consider features of fuzzy summaries to point out a set of interesting properties. The final goal of this work is to offer to the end-user a set of high level operators for exploring and querying summaries in decision making environments.

The paper is organized as follows: section 2 presents multidimensional databases and OLAP operators. Section 3 presents an overview of the SaintEtiQ system. Our contribution consists in section 4 and section 5. Indeed, section 4 describes the data model we designed over the summaries to support decision making processes. In section 5 we define a collection of unary operators to handle summaries. And finally in section 6 we conclude and propose future works.

2 Multidimensional Databases

Multidimensional database technology is a key factor in the interactive analysis of large amounts of data for decision-making purposes.

On-Line Analytical Process framework is very interesting for data analysis since it handles hierarchies to represent data at different levels of granularity. It consists in a set of technologies to collect, store, and treat multidimensional data for analysis. With OLAP tools, users do not aim at manipulating data at individual level. They rather want to see data at aggregated, consolidated level for the discovery of trends. The datacube was introduced by Gray et al. in [8]. A cube is a set of data organized according to dimensions. A *measure* is the value contained in a cell, associated to the values taken on *dimensions* composing the cube.

OLAP operators. Here is a partial list of transactional and granularity operators used to manipulate data cubes [1]:

Transactional manipulation, considers the extension of well-known operations of the standard relational framework. It deals with multidimensional information.

- The *Slice* operation consists in selecting slices from the cube by using a criterion on a dimension.

- The *Dice* operation consists in selecting a subset of cells matching a criterion on the measure.

- The *Projection* operation consists in deleting one or several dimensions.

- The *Merge* operation consists in merging cube dimensions according to aggregation function.

Granularity operators, define navigation operations through hierarchies constructed on dimensions.

- The *Roll-up* operation consists in describing the cell values at some higher level of granularity on a dimension.

The above enumeration concerns only unary operators. Obviously, there exist a few more n-ary operators in OLAP algebra such as cartesian product, join or drill-down They are out of the scope of this communication, but we have to study them in the future.

3 Overview of SaintEtiQ system

The SaintEtiQ approach considers a primary relation $R(A_1, \dots, A_n)$ in the relational database model, and constructs a new relation $R^*(A_1, \dots, A_n)$, in which tuples z are summaries and attribute values are fuzzy linguistic labels [14] describing a sub-table of R .

The preprocessing. All tuples t of R are rewritten using background knowledge (BK), such as fuzzy linguistic labels. For instance, a fuzzy linguistic label *young*, gather several values $t.AGE$ of distinct tuples t on the attribute *AGE*. Consider the tuple $Burns = \langle NPPBoss, 87000\$ \rangle$ for $(name(id) = \langle occupation, income \rangle)$ from relation about SIMPSONS-CHARACTERS. The translation step converts this tuple into two user-defined vocabulary tuples:

$Burns [1] = \langle businessman, enormous \rangle$

$Burns [2] = \langle firmmanager, enormous \rangle$

An example is given in table 1, where ϕ is the appropriate satisfaction degree obtained from the membership grade of each record to linguistic terms.

This translation step support the process of

NAME (id)	OCCUPATION ϕ	INCOME ϕ
<i>Apu</i> [1]	businessman .7	miserable 1.0
<i>Apu</i> [2]	shopkeeper 1.0	miserable 1.0
<i>Burns</i> [1]	businessman .9	enormous 1.0
<i>Burns</i> [2]	f.manager 1.0	enormous 1.0
<i>Moe</i> [1]	businessman .7	miserable 1.0
..

Table 1: Example of translation step of the relation SIMPSONS-CHARACTERS

finding the best representation of a database tuple according to BK provided by the user. It consists in transforming the raw data into a new tuple by replacing the original value by the set of descriptors defined in the BK. The result of translation is considered as the first level of summarization. $Burns [k]$ is in fact the intentional description of a summary of database records close to *Burns*.

The summary. Each summary z provides a synthetic view of a part of the database. The subset of database records R_z involved into

the summarization is usually called the *extent*, whereas the summarized description I_z of these database records is the *intent*. For convenience, we will denote by z the *intent* I_z of summary z .

Intent: The intentional description $z = \langle z.A_1, \dots, z.A_n \rangle$ of a summary describes similar features of tuples in R_z . It allows to generalize the descriptions of database tuples, attribute by attribute. Each $z.A_i$ is a fuzzy set represented by some linguistic label d . A descriptor d generalizes the attribute values of database records in the *extent* of z . Label d is associated with a weight α corresponding to the highest membership grade of the $t.A_i$'s to d : $\alpha = \max_{t \in R_z} \{d(t.A)\}$, where $d(t.A) = \phi_d(t.A)$. This α_d corresponds to the satisfaction of having d as a summary descriptor. $z = \langle z.A_1, z.A_2, \dots, z.A_n \rangle$ with $z.A_i \in \mathbf{F}(D_A^+)$, $1 \leq i \leq n$, where $\mathbf{F}(X)$ denotes the set of fuzzy sets on X . D_A^+ is the translated attribute domain of A defined as the finite set of linguistic terms describing the attribute A . For instance, $D_{INCOME}^+ = \{miserable, modest, reasonable, enormous, \dots\}$

Extent: A summary is defined in an extensional manner with a collection of records $R_z = \{t_1, t_2, \dots, t_n\}$, where t_i is a database tuple. For instance, consider $R_z = \{Apu, Burns, Moe\}$. The rewritten tuples identified by $t[1]$, $t \in R_z$, in table 1 allow to define the summary z . The intentional description of z is given by :

$$z = \langle \{.7/businessman + .8/artist\}, \{1.0/enormous + 1.0/miserable\} \rangle$$

Cardinality: The cardinality of z is defined as $card(R_z) = \sum_{t \in R_z} w(t)$, where $w(t)$ is the weight associated to the tuple t depending on the number of rewritten tuples. For instance, consider record t has 2 distinct rewritten tuples: $\langle artist, miserable \rangle$ and $\langle artist, enormous \rangle$ then $w(t) = 1/2$. $card(R_z)$ corresponds to the representativity of the summary z according to the database R . The descriptor cardinality $card_{z.A}(d)$ relative to the extent of z determines the proportion of database records involved into the generalized description of R_z with the linguistic label d .

4 Data Model

4.1 Summary and space partition

We want to build a generic materialized view from all or parts of the relation R . For this, we propose to define the notion of *summary partition* as a collection of summaries verifying a constraint.

Definition 1 (Summary Partition). *Let $\mathcal{A} = \{A_1, \dots, A_m\}$ be a set of attributes. A partition P is a set of summaries z built on \mathcal{A} and satisfying the weak orthogonality property. $\mathbf{P} = \{z_0, \dots, z_n\}$ such as : $\forall i \neq j \in [0..n], \exists k \leq m, z_i.A_k \cap z_j.A_k = \emptyset$ where m is the number of attributes.*

The property of weak orthogonality means that we can't have two summaries in the same summary partition that share the same intentional representation on each attribute. We say that two summaries are *conflictual* if they do not verify the property of weak orthogonality.

Definition 2 (Partition space). *Let $\mathcal{A} = \{A_1, \dots, A_k\}$ be a set of attributes. The partition space noted \mathcal{P} is the set of summary partitions built on X , $\forall X \subseteq \mathcal{A}$.*

We note that this data model of fuzzy summaries is based on multidimensional vague concepts. It performs a conjunctive approach on each attribute, whereas possibilistic databases consider weighted disjunctive attribute values. It is worth to mention that the associated semantics of algebraic operators defined over these two models is not similar, e.g. attribute *AGE* can be represented by old *and* young people in the first model, and by young *or* old people in possibilistic databases.

4.2 Hierarchy

The SaintEtiQ system [11] incrementally builds a summary hierarchy noted by H_R as shown by figure 1. Nodes of this rooted tree are summaries defined at different abstraction levels. The root is the more general summary whereas leaves are the most specific ones.

The partial ordering on summaries: Given z and z' , two elements of the summary hierarchy, we define the relation \preceq as : $(z \preceq z') \Leftrightarrow (z.A_i \subseteq z'.A_i, \forall A_i \in \mathcal{A})$. If $z \preceq z'$ we say that z' gen-

eralizes z , such that the linguistic descriptors of z' on each attribute are either at least as general as those of z , or they are new ones representing another trend of the group of tuples in $R_{z'} - R_z$.

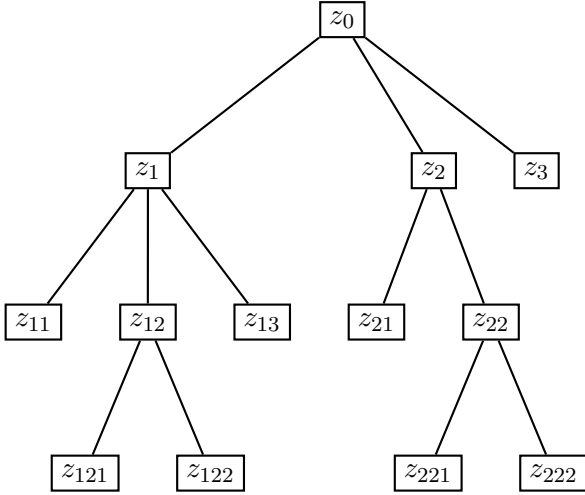


Figure 1: Example of SaintEtiQ hierarchy

Definition 3 (Summary height). *In our summary tree the height $h(z)$ of each node is the longer path from this node to one of its leaves, e.g. if z is a leaf then $h(z) = 0$.*

Abstraction level. An abstraction level is regarded as a cut of a summary tree.

Definition 4 (Cut). *A cut C of a summary tree is a set of summaries verifying the two following properties : (1) the weak orthogonality and, (2) the completeness: $\bigcup_C(R_z) = R$. We note $\mathcal{C}(H_R)$ the set of cuts built on the hierarchy H_R .*

$\mathcal{C}(H_R)$ is a subset of the partition space \mathcal{P} (see def 2): $\mathcal{C}(H_R) \subset \mathcal{P}$.

Example For the hierarchy in the figure 1, the set of cuts is:

$$\mathcal{C}(R): \left\{ \begin{array}{l} P_0 = \{z_0\} \\ P_1 = \{z_1, z_2, z_3\} \\ P_2 = \{z_{11}, z_{12}, z_{13}, z_2, z_3\} \\ P_3 = \{z_{11}, z_{121}, z_{122}, z_{13}, z_2, z_3\} \\ P_4 = \{z_1, z_{21}, z_{22}, z_3\} \\ P_5 = \{z_1, z_{21}, z_{221}, z_{222}, z_3\} \\ P_6 = \{z_{11}, z_{12}, z_{13}, z_{21}, z_{22}, z_3\} \\ P_7 = \{z_{11}, z_{12}, z_{13}, z_{21}, z_{221}, z_{222}, z_3\} \\ P_8 = \{z_{11}, z_{121}, z_{122}, z_{13}, z_{21}, z_{22}, z_3\} \\ P_9 = \{z_{11}, z_{121}, z_{122}, z_{13}, z_{21}, z_{221}, \\ z_{222}, z_3\} \end{array} \right.$$

	OCCUPATION	α_d	INCOME	α_d
z_{11}	businessman	0.7	miserable	1.0
	artist	0.8	enormous	1.0
z_{121}	shopkeeper	0.9	miserable	1.0
z_{122}	sch.qual.employee	0.3	modest	1.0
z_{13}	sch.qual.employee	0.8	enormous	1.0
z_2	artist	0.8	modest	1.0
			miserable	1.0
z_3	businessman	0.8	enormous	1.0

Table 2: Presentation of partition P_3

The SaintEtiQ hierarchy will help us to define an algebra for on-line analysis and high-level presentation of data. It will also guide us in the exploration of the summaries providing a relevant starting point for analysing process, as well as a backbone for granularity operators like roll-up and drill down.

The algebra we will define over our data model is close on the set of partition space \mathcal{P} . It means that the result of each operator over a summary partition P gives a summary partition P' , i.e. $(P, P') \in \mathcal{P}^2$.

5 Operators

In this section we focus on definitions of unary operators. Binary operators are out of the scope of this paper but they have to be studied in order to deal efficiently with fuzzy summaries. In the following we present the selection, roll-up, merge and projection, which are the main part of our core algebra over summary partitions.

5.1 SELECTION

Definition 5 (Selection). *The selection operator σ on the summary partition P from R , gives a new summary partition P' such that: $P' = \sigma(P, pred(z))$, where $pred(z)$ is the selection predicate on:*

- Summary properties (Dice), or
- Attribute values (Slice).

The selection operation allows us to extract a sub-set of the summaries of a partition without any update of their intentional and extensional

descriptions.

Dice. The first type of query is the selection from summary properties. Based on the hierarchical organization of the summaries we can express any structural constraint, like the granularity using the height of summary. The predicate is then $h(z) = n, n \in \mathbb{N}$. For instance, if we want to retrieve only the leaves from the hierarchy (figure 1), $pred(z)$ is defined as $h(z) = 0$, gives the partition $P' = \{z_{11}, z_{121}, z_{122}, z_{13}, z_{21}, z_{221}, z_{222}, z_3\}$.

Slice. The second type of query is the selection over fuzzy attribute values of summaries and their membership grades. We define it as: $z.A \theta \tilde{v}$, where θ is a fuzzy set operator such as $=_F, \subseteq_F, \cap_F, \subset_F$, etc, and $\tilde{v} \in F(D_A^+)$ is a fuzzy set. For instance, $P' = \sigma(P, z.INCOME \subseteq_F \{1.0/enormous, 1.0/modest\})$. Given the partition P_3 of table 2, $P' = \{z_{11}, z_{122}, z_{13}, z_2, z_3\}$.

5.2 ROLL-UP

The set of all the partitions is partially ordered, thanks to the relationship defined over the summaries into the SaintEtiQ hierarchy.

Definition 6 (Roll-up). *The roll-up operation is a generalization of the partition at a higher level of granularity:*

$$P' = Roll-up(P = \{z_1, \dots, z_i, z_{i+1}, \dots, z_n\}, \{z_1, \dots, z_i\})$$

$$P' = \{z', z_{i+1}, \dots, z_n\}$$

such that $\forall z \in \{z_1, \dots, z_i\}, z \preceq z'_i$ and $R_{z'} = \bigcup_P (R_{z_1} \dots R_{z_i})$

Example. In table 2, $Roll-up(P_3, \{z_{121}, z_{122}\}) = P_2$, with $z_{121} \preceq z_{12}$ and $z_{122} \preceq z_{12}$ and $R_{z_{12}} = R_{z_{121}} \cup R_{z_{122}}$.

5.3 MERGE

The merge operation needs a summary partition and an aggregation function. Merging summaries of a partition consists in aggregating membership grades of similar descriptors over each attribute.

Definition 7 (Merge). *Let $P = \{z_1, \dots, z_k, \dots, z_n\}$ be a summary partition. We then define the merge*

operator as:

$$P' = merge(P, \{z_1, \dots, z_k\}).$$

$$P' = \{z^*, z_{k+1}, \dots, z_n\}$$

For instance, $merge(P, \{z_i, z_j\})$ with $z_i = \langle \{\alpha_a/d_1\}, \{\alpha_b/d_2\} \rangle$ and $z_j = \langle \{\alpha_c/d_1\}, \{\alpha_d/d_3\} \rangle$ gives birth to $z^ = \langle \{max\{\alpha_a, \alpha_c\}/d_1\}, \{\alpha_b/d_2\}, \{\alpha_d/d_3\} \rangle$*

The usual aggregating function is the maximum. We suppose that only merge operations such as summary z^* is not conflictual with other summaries $z \in P'$ ($z \neq z^*$) are allowed. About the extension corresponding to the merge operation we can say that:

$$R_{z^*} = R_{z_1} \cup \dots \cup R_{z_k} \quad \text{and,}$$

$$card(z^*) = card(z_1) + \dots + card(z_k)$$

	OCCUPATION	α_d	INCOME	α_d
z_{13}	sch.qual.employee	0.8	modest	1.0
z_{122}	sch.qual.employee	0.3	miserable	1.0
merge($P_3, \{z_{13}, z_{122}\}$)				
z^*	sch.qual.employee	0.8	miserable	1.0
			modest	1.0

Table 3: Merging on the partition P_3

The table 3 shows how we can merge two summaries z_{13} and z_{122} .

5.4 PROJECTION

We consider here the operator consisting in reducing the number of attributes describing the summaries of a partition.

Definition 8 (Projection). *Let P be a summary partition. We define the projection operator as:*

$$P' = \Pi_{A_1, \dots, A_k}(P)$$

$$P' = \{z' / \exists z \in P \wedge z' = \pi_{A_1, \dots, A_k}(z)\}$$

where the summary projection $\pi_x \langle x, y \rangle = \langle x \rangle$.

The extensional descriptions of summaries remain the same after applying the projection operator. As for the merge operation the projection operation is able to generate partitions which do not verify the weak orthogonality property. For instance, the possible projections of the two summaries in table 4 have to involve the A attribute.

Indeed, $\pi_{B,C}(z_1, z_2)$, $\pi_B(z_1, z_2)$, $\pi_C(z_1, z_2)$ lead to conflictual summaries, i.e. with indistinguishable descriptions.

	A	B	C
z_1	a_1	b_1	c_1
z_2	a_3	b_1	c_1

Table 4: Example of representation on attributes A, B and C

6 Conclusion

In this paper we first introduced the concept of summary partition. Beforehand summaries are organized into a hierarchy. We then defined unary operators over the SaintEtiQ-oriented partitions. Current and future works concern the definition of binary and higher-level operators. This work is intended to provide the core algebra of an effective and rich tool for visualizing, querying and accessing the data through the summaries as OLAP datacubes operators do.

References

[1] R. Agrawal, A. Gupta, and S. Sarawagi. Modeling multidimensional databases. *in Proc. of ICDE*, 1997.

[2] I. Blanco, D. Sánchez, J.-M. Serrano, and M. A. V. Miranda. A new proposal of aggregation functions: The linguistic summary. In *IFSA*, pages 127–134, 2003.

[3] Patrick Bosc, Olivier Pivert, and Laurent Ughetto. On data summaries based on gradual rules. In *Fuzzy Days*, pages 512–521, 1999.

[4] Juan C. Cubero, Juan Miguel Medina, Olga Pons, and María Amparo Vila Miranda. Data summarization in relational databases through fuzzy dependencies. *Inf. Sci.*, 121(3-4):233–270, 1999.

[5] S. Babu, G. Minos, and R. Rajeev. SPARTAN: A model-based semantic compression system for massive data tables. In *Proc. of*

the 2001 ACM Intl. Conf. on Management of Data (SIGMOD 2001), pages 283–295, May 2001.

- [6] M. Delgado, C. Molina, D. Sánchez, L. R. Ariza, and M. A. V. Miranda. A flexible approach to the multidimensional model: The fuzzy datacube. In *CAEPIA*, pages 26–36, 2003.
- [7] Janusz Kacprzyk and Slawomir Zadronzny. On interactive linguistic summarization of databases via a fuzzy-logic-based querying add-on to microsoft accessè. In *Fuzzy Days*, pages 462–472, 1999.
- [8] Jim Gray, Surajit Chaudhuri, Adam Bosworth, Andrew Layman, Don Reichart, Murali Venkatrao, Frank Pellow, and Hamid Pirahesh. Data cube: A relational aggregation operator generalizing group-by, cross-tab, and sub-totals. *Data Min. Knowl. Discov.*, 1(1):29–53, 1997.
- [9] A. Laurent. Querying fuzzy multidimensional databases: unary operators and their properties. *Int. J. Uncertain. Fuzziness Knowl.-Based Syst.*, 11(Supplement):31–45, 2003.
- [10] Henri Prade and Claudette Testemale. Generalizing database relational algebra for the treatment of incomplete/uncertain information and vague queries. *Inf. Sci.*, 34(2):115–143, 1984.
- [11] G. Raschia and N. Mouaddib. A fuzzy set-based approach to database summarization. *Int. Journal of Fuzzy Sets and Systems*, 129(2):137–162, July 2002.
- [12] Dan Rasmussen and Ronald R. Yager. Summary sql - a fuzzy tool for data mining. *Intell. Data Anal.*, 1(1-4):49–58, 1997.
- [13] R.R. Yager. A new approach to the summarization of data. *Information Sciences*, 28(1):69–86, October 1982.
- [14] Lotfi A. Zadeh. Fuzzy sets. *Information and Control*, 8(3):338–353, 1965.