Proceedings of the *International Conference on Theory and Application of Mathematics and Informatics ICTAMI 2005* - Alba Iulia, Romania

A COMBINATORIAL TOPOLOGY APPROACH OF DATA CONSOLIDATION

Cristian Kevorchian and Laurenţiu Modan

Abstract.Data consolidation is the process synthesizing pieces of information into a essential knowledge single block. The highest level in data consolidation process is referred through the data dimension concept. There is a number of different dimensions from which a given pool of data, can be analyzed.Multidimensional conceptual view will be the way for most business persons to organize their global enterprize. An user-analyst's view of the global enterprize's universe is a multidimensional one. Accordingly with this way, the user analyst's conceptual view of OnLine Analytical Processing(OLAP) models, introduced in 1993, is a multidimensional one. Multidimensional data type implemented through a topological model, provides an important theoretic tool for OLAP[9]. The main target of data management is to retrieve data, following a query pattern. This query pattern can be provided as a restriction of the class of simplicial complexes provided by spatial data objects.Following this strategy we can design an ameliorate tool to retrieve, data from a multidimensional data structure.

2000 *Mathematics Subject Classification*: 68P20, 55U15.

1. Combinatorial model

Combinatorial topology provides us complex methods for point sets formal described[1]. Let a finite set $\mathcal{V}$=$\{v^i, i = 1..k\}$ be and a collection $\mathcal{K}$:

$$\mathcal{K}=\{v^{11}...v^{1i_1}, v^{21}...v^{2i_2}, ..., v^{k1}...v^{ki_k}\}, \text{ where } i_1, ...i_k \in \{1..k\}$$

a family with k elementary data. We will consider every (p+1) elementary data , as a p-simplex, σ_p.

Let $\prec$ be a partial ordering relation on $\mathcal{K}$. We will note $\sigma_p \prec \sigma_q$ when σ_p is a subsequent of σ_q and we'll say that $(\mathcal{K}, \prec)$ is a simplicial complex([4][5]) if and only if :
a) each elementary data $[v^i]$ is a element of the collection $\mathcal{K}$ as a σ_0;
b) if $\sigma_p \in \mathcal{K}$ and $\sigma_p \prec \sigma_q$ then $\sigma_q \in \mathcal{K}$.
The dimension of $\mathcal{K}$ is given by the largest dimension of its simplex and it will be noted by $dim(\mathcal{K})$. The dynamic data structure (queries family) is based on the simplex connection, named

chain connection.

Our approach provides a geometrical representation of the operations on data family ,$\mathcal{K}$ in terms of connected convex polyhedra. We will try to implement a computational process on the data family. Using this procedure we will obtain a new data family. Now we'll analyze the simplicial complex connectivity[2][9]. Let two simplex σ_p,$\sigma_q \in \mathcal{K}$ be. We will say that our simplexes are linked through a chain if there is a simplex sequence:

$$\sigma_{\alpha_1}, \sigma_{\alpha_2}, ..., \sigma_{\alpha_h}$$

a) σ_{α_1} is a face of σ_p;
b) σ_{α_h} is a face of σ_q;
c) σ_{α_i} and σ_{α_i+1} has a common face σ_{p_i} , for $i = 1, 2, ..., (h-1)$
We will say that such chain of connection has (h-1)length and we also say that the chain is a q-connected if:

$$q = min(\alpha_1, \beta_1, \beta_2, ..., \beta_{h-1}, \alpha_h)$$

where β_j, for $j = 1, ..., h-1$ are dimensions for the intermediates simplexes. Let γ_q be a relation on a simplicial complex $\mathcal{K}$ of the type:

_ is q-connected to _

It is easy to see that:

1) if $\sigma_p \in \mathcal{K}$ then $(\sigma_p, \sigma_p) \in \gamma_q$;

2) if $(\sigma_l, \sigma_p) \in \gamma_q$ then $(\sigma_p, \sigma_l) \in \gamma_q$;

3) if $(\sigma_l, \sigma_m) \in \gamma_q$ and $(\sigma_m, \sigma_p) \in \gamma_q$ then $(\sigma_l, \sigma_p) \in \gamma_q$.

Therefore γ_q is an equivalence relation and we will note by Q_q the cardinally of the set $\mathcal{K}/\gamma_q$

2.Data cube and data consolidation

When we use any OLAP product,we use data aggregation and we load large queries providing elements for data warehouse and online transaction processing (OLTP)[2]. Calculating and storing data aggregation in a separate area for later retrieval by the OLAP product allows to the users to retrieve a large amount of information while this reduces the amount of data to be processed. More this functionality, which relieves data warehouses of the large workload potentially for aggregating and responding to such queries by users, gives us its own special set of issues. Two of the most common issues, data explosion and data sparsity, are caused by OLAP products' need to establish and access aggregation storage. Although Analysis Services handles efficiently both issues, understanding them is equivalently to improve your cube design and performance. Let be the following, very simple relational data structures R_1 and R_2 with:

R_1	$k1$	$v1$	$v2$
1	100	3	$Null$
2	101	2	1

and

R_2	$k2$	$v3$	$v4$	$v5$
1	200	1	$Null$	$Null$
2	201	$Null$	3	1

where $k1$ and $k2$ are structure instances keys. Set also be the "cub" operator which is an extension of "GROUP BY" clause which generates subtotals for all permutations realized with grouping columns.

Cub	$k1$	$k2$	$v1$	$v2$	$v3$	$v4$	$v5$
1	100	200	3	$Null$	1	$Null$	$Null$
2	100	201	3	$Null$	$Null$	3	1
3	101	200	2	1	1	$Null$	$Null$
4	101	201	2	1	$Null$	3	1

Let $\mathcal{Q}$=$\{q_1, q_2, ..., q_5\}$, be a query family on the data cub:
1. q_1: $k2 \leq 200 \vee v1 \geq 3$
2. q_2: $k2 \geq 200 \vee v2 \neq NULL$
3. q_3: $v3 \neq NULL$
4. q_4: $k1 \geq 101 \vee v4 \neq NULL$
5. q_5: $k2 - k1 > 100 \wedge v5 = NULL$
We'll establish a geometrical approach of the query family cube through an associating with a simplicial complex, where every simplex will show the existence of a record queries contained in q_1,q_2,...,q_5, we will show in section 4. Through q-analyze we will try to obtain a global indicator of data cub query process.(see [7][2][1]).

3.Q-analysis over OLAP

Through Q-analyze of the complex attached to the data cube we can obtain an array (Q_q), $q \in \{0, 1, .., dim(\mathcal{K})\}$, where Q_q is the number of q-connected data components. The array Q can be considered as a *global indicator* of the simplicial complex K, attached to the data cub but this didn't provide a unique characterization of the attached simplicial complex to the instance family of our data structures. Two different instances can be characterized by same Q_q array. More details about this analysis can be obtained through the study of chain connectivity, given by:

$$\Psi(\mathcal{K}) = \frac{2}{(n+1)(n+2)} \sum_{q=0}^{n} (q+1)Q_q, \text{ where } n = dim(\mathcal{K}).$$

Let $\mathcal{K}_1$ and $\mathcal{K}_2$ be two data structures. We have :

$$\mathcal{K}_1 \equiv \mathcal{K}_2 \text{ iff } \Psi(\mathcal{K}_1) = \Psi(\mathcal{K}_2)$$

So, the results of Q-analysis for two different instances of data structure are the same, iff the numerical evaluation of $\Psi(\mathcal{K}_1)$ and $\Psi(\mathcal{K}_2)$ is the same. The local properties of the individual simplex is very important in the analysis of data's structure instances. The measure of this indicator is *eccentricity of simplex*[see 5]:

$$ecc(\sigma) = \frac{\hat{q} - \overline{q}}{\hat{q} - 1},$$

where $\hat{q}$ is the diagonal element of the row corresponding to σ in the "shared face matrix"([8][6]), given by SF=$\Lambda\Lambda^T - \Omega$,when Λ is the incidence matrix, Ω is a unit matrix and $\overline{q}$ is the largest non diagonal value in a σ entrance. The eccentricity, attached to a simplex, is infinite iff this simplex is disconnected from all other simplex of the complex.

Each $Q_q, q \in \{0, 1, ..., dim(\mathcal{K})\}$ belongs to the first structural vector Q, namely a vector, having q-connected components number in the instance data structure family $\mathcal{K}$. Each of such components can belong to several simplexes. The vector Q didn't focus on this issue. We can use as $\overline{Q}$ a *second structural vector of the data structure* defined through :

$$\overline{Q}_q = 1 - \tfrac{Q_q}{n_q}$$

where n_q is the total number of d-simplexes($d \geq q$) belongs to Q_q as connected components, from the dimensional level q. Vector $\overline{Q}$ can be interpreted as a modified measure of data connectivity degree at the level q,where $q \in \{0, 1, ..., dim(\mathcal{K})\}$ and the improve data components analysis. The ratio $\frac{Q_q}{n_q}$ defines the number of the connected components per one simplex σ with , $dim(\sigma) \geq q$. We can obtain a new form of the $\mathcal{K}$ family, where $Q_0 = k$. We'll identify the data structure instances family, with the component $Q_0 = 1$ as a simplicial complex.Let $K_Y(X, \lambda)$ the attached complex with $dim(K) = n$. This theoretical environment leads us to design a data transformation procedure[4]. The algebraic transformation of our data structure can be written as a family of rules[4]. All this rules will be implemented in specialized language, by examples MGS(***M***odèle ***G***énerale de ***S***imulation(de système dinamique))[4]. An important issue is to obtain similar tools for data analysis in Bussiness Inteligence(BI) applications[4].

4.The model for data structure analysis

Let be the following complex attached to data cube defined at Section 2 :

$$K = \{x_1x_3, x_1x_2x_4, x_1x_2x_3, x_4x_5, x_2x_4x_5\}$$

Let the relation λ, with $\lambda \subseteq Y \times X$ where $X = \{x_1, x_2, ..., x_5\}$ and $Y = \{y_1, y_2, ..., y_4\}$ where $y_1, y_2, ..., y_4$ are simplexes generated by the data cube dimensions. $K_X(Y; \lambda^{-1})$ is the simplicial complex provided by the relation λ^{-1} attached to cube data slice. $\Lambda = (\lambda_{i,j})$ is the incidence matrix, with 4 lines and 5 columns:

λ	x_1	x_2	x_3	x_4	x_5
y_1	1	0	1	0	0
y_2	1	1	0	1	0
y_3	1	1	1	1	1
y_4	0	1	0	1	1

$K_Y(X, \lambda)$ complex is given by the evaluation of the "shared face matrix", $\Lambda\Lambda^T - \Omega$,where Ω is 4×4 1's matrix of type :

$$\begin{array}{ccccc}
y_1 & y_2 & y_3 & y_4 & \\
1 & 0 & 1 & -1 & y_1 \\
 & 2 & 2 & 1 & y_2 \\
 & & 4 & 2 & y_3 \\
 & & & 2 & y_4
\end{array}$$

Therefore we have the following structure vector for $K_Y(X, \lambda)$:

$$\text{Q}=(\overset{4}{1}, \overset{3}{1}, \overset{2}{1}, \overset{1}{1}, \overset{0}{1})$$

and the analysis result is a triangular matrix array. The -1 value means that y_1 and y_4 are not connected.

q=4	$q_4 = 1$	y_3
q=3	$q_3 = 1$	y_3
q=2	$q_2 = 3$	y_3, y_2, y_4
q=1	$q_1 = 4$	y_3, y_2, y_4, y_1
q=0	$q_0 = 1$	all

Based on Q-analysis, we can provide a new approach of the data cube as following structure :

$$\overline{\mathcal{M}} = \{y_3, y_3, y_3y_2y_4, y_1y_2y_3y_4, y_1y_2y_3y_4\}$$

From this evaluating of the above matrix $\Lambda^T\Lambda - \Omega$ we obtain the pattern for the complex $K_X(Y, \lambda^{-1})$ where

$$Q = (\overset{2}{2}, \overset{1}{1}, \overset{0}{1})$$

$$\begin{array}{cccccc}
x_1 & x_2 & x_3 & x_4 & x_5 & \\
2 & 1 & 1 & 1 & 0 & x_1 \\
 & 2 & 0 & 2 & 1 & x_2 \\
 & & 1 & 0 & 0 & x_3 \\
 & & & 2 & 1 & x_4 \\
 & & & & 1 & x_5
\end{array}$$

and for $K_X(Y, \lambda^{-1})$ we obtain:

q=2	$Q_2 = 2$	x_1, x_2, x_4
q=1	$Q_1 = 1$	all
q=0	$Q_0 = 1$	all

So, the data structure obtained is:

$$\widetilde{\mathcal{K}} = \{x_1x_2x_4, x_1x_2x_3x_4x_5, x_1x_2x_3x_4x_5\}$$

The components of the second structural vector are:

$$\begin{array}{ll}
q = 2 & \overline{Q}_2 = 1 - (\frac{Q_2}{3}) = 0.66 \\
q = 1 & \overline{Q}_1 = 1 - (\frac{Q_1}{6}) = 0.17 \\
q = 0 & \overline{Q}_0 = 1 - (\frac{Q_0}{6}) = 0.17
\end{array}$$

Practically, the components of $\overline{Q}$ show us the connectivity degree of the data cube.

5.Conclusion

The main goal of the combinatorial topology data structures study is to provide a data organization measure for optimizing data consolidation process. Business intelligence(BI) applications design involves the large amount of data processing. The results' quality of BI applications depends on the quality of data representation, reflected in the consolidation process. We can say that our approach provides a measure of data structuring quality. It is very important in designing process economy to have a indicator for data organization quality.

References

[1] Atkin, R.H. (1977) Combinatorial connectivities in social systems. An application of simplicial complex structures to the study of large organizations, Birkhauser Verlag, 1977.

[2] Collosi N., Malloy W., Reinwald B., *Relational extensions for OLAP* IBM System Journal vol. 41,no.4,2002,pag. 714-731.

[3] Giavitto Jean-Louis, Olivier Michel, *The Topological Structures of Membrane Computing* Rapport de Recherche no. 70/2001, CNRS-Univeriste d'Evry Val d'Essonne.

[4] Kevorchian H. C.,*Semiotics and artificial intelligence an algebraic topological approach* R.R.L. XXIII,Ed.Academiei 1986.

[5] Konstantin Y. Degtiarev, *System analysis:mathematical modeling and approach to structural complexity measure using polyhedral dynamics approach*, Complexity International, vol.7,2000,http://www.csu.edu.au/ci/vol07/degtia01/.

[6] Sonis Michael, Hewings J.D., *Introduction to Input-Output Structural Q-analysis, REAL 00-T-1,http://www.uiuc.edu/unit/real.*

[7] Teruaki Aizawa, *Some.Homology Theoretic Structures of Languages* Systems,Computers,Controls: vol. 2,no. 5,1971.

[8] Valencia E. , *Un Modèle Topologique pour le Raisonnement Diagrammatique* Mèmoire, DEA de Science Cognitives, LIMSI, Université Paris-XI, Orsay ,Avril-Août,1997.

[9] Witten I. H., Eibe F., *Data Mining*, Morgan Kaufmann Publishers,2000.

Cristian Kevorchian
Department of Computer Science
University of Craiova, Mathematics-Informatics Faculty
Address: A.I. Cuza, 13, Craiova,1100
e-mail:*cristian.kevorchian@gmail.com, www.ckro.go.ro*

Laurenţiu Modan
Department of Mathematics
Computer Science Faculty
Address: Calea Dorobanţilor 15-17, Sector 1, Bucharest
e-mail:*modan_laurant@yahoo.fr*