ATLANTIS PRESS

Advances in Intelligent Systems Research, volume 133

2nd International Conference on Artificial Intelligence and Industrial Engineering (AIIE2016)

Technology of Exploratory OLAP Based on the Integral Analytical Model

Anna Korobko

Institute of Computational Modelling of SB RAS, Krasnoyarsk, Russia

***Abstract*—The paper presents the original approach to the exploratory OLAP of the heterogeneous resources based on the integral analytical model. Applying the multidimensional data model as a technological fundament allows to analyze the structure of the joining sources according to unified notation, allows to present information in form of native multidimensional view and provides efficient integration of the sources by matching the dimensions. Storing information about the physical location and comparability of the analytical concepts within multidimensional model enables to form automatically the exploratory queries. The proposed approach leans on analysis and matching resources on the conceptual level (i.e. without physical data consolidation) that allow to process data with a lot of records and entities. In the paper, the requirements to exploratory OLAP technology are defined and the main issues and their solutions are formulated based on preliminary scientific results. The principal stages for analytical exploration support according to author's method of the integral OLAP-modeling are described.**

Keywords-exploratory OLAP; heterogeneous data; big data; conceptual analytical model; integral OLAP-model

I. INTRODUCTION

The modern level of computer science technologies provides the basis for forming great heterogeneous data space. Comprehensive on-line data analysis from miscella-neous sources within the domain and its outside will allow to improve the efficiency of the management decisions significantly. One of the most popular and effective technology of data processing is OLAP (on-line analytical processing) that gives the opportunity for intuitive manipulation of multidimensional data and execution of the analytical queries "on the fly" [1, 2]. Therefore the topical problem is development of OLAP technology in a part of exploratory analytical processing of the heterogeneous data from different sources. However, it requires a powerful technological base that provides the full-cycle support of all stages of native analytical data processing.

The key issue that should be solved is identification of the data model which can join the heterogeneous resources. Apart from this, the important issue is high speed of integrating the new sources into the common model because the analytical exploring is related with execution of complex technological operations and huge amount of the data should be processed. A lot of analytical concepts of the common model and their variety require the special approach to manipulation of all joined data.

Taking into account the defined above issues, this paper presents an original approach to the exploratory OLAP [3] technology based on hybridization of the following solutions: OLAP, multidimensional model, formal conceptual analysis and lattice theory. Section 2 presents the formalization of the principal stages for analytical exploration support according to author's method of the integral OLAP-modeling. Sections 3-7 describe each of the stages in details.

II. THE EXPLORATORY OLAP BASED ON THE INTEGRAL OLAP-MODEL

The main idea of propose approach is to form the data source structure by analyzing only their metadata without complete data downloading, to discover the common analytical concepts and to construct the integral model that will provide intuitive manipulation of the joined resources. Figure I shows the decomposition of integrating the new source to the common analytical model. Diagram illustrates the stages sequences of analytical exploration support, where each stage has input and output parameters in accordance with data processing methods.

The stage "Construct the interrelated multidimensional model of the new source" is aimed at standardization of the joining data sources by constructing the multidimensional model of the source according to unified notation regardless of the initial storage format. Original edition to multidimensional data model theory is definition of all possible analytical relations between dimensions and measures taking into consideration the data source consistency. The interrelated multidimensional model (IMM) is a triple (D, F, R) where D is a set of dimensions, F is a set of measures, R is a relation of comparability between elements from D and F. On the one hand, this model allows to present information in form of native multidimensional view, on the other hand, it provides support for automatic query generation based on the set of all available analytical concepts. One of the most critical stages is "Integrate the source interrelated multidimensional model into the common interrelated multidimensional model". This stage is aimed at matching the interrelated multidimensional models of the joining sources in order to define the common analytical concepts.

Author's method of the integral OLAP-modeling [4] is based on analysis of context K – binary matrix where the rows are correspond to measures of the multidimensional model, the columns are correspond to dimensions and the matrix elements show the defined relation of comparability between them. In general case, the common interrelated multidimensional model is a graph that can be presented as a reduced adjacency matrix

– the context. This process is carried out at the stage "Form matrix context based on the common interrelated multidimensional model".

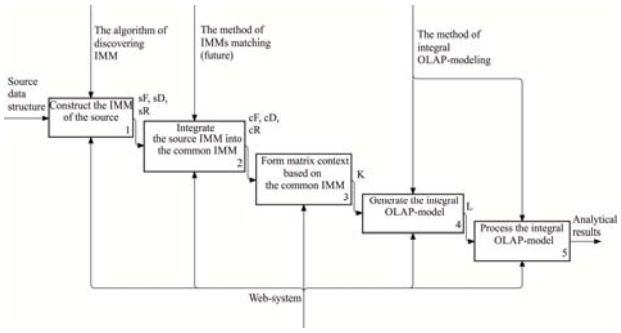

FIGURE I. IDEF0 DIAGRAM OF EXPLORATORY OLAP BASED ON THE INTEGRAL OLAP-MODEL

The stage "Generate the integral OLAP-model" is aimed at the context clustering – generation of the subsets of K. These subsets are characterized by extent (i.e. the set of dimensions) and intent (i.e. the set of measures) and present the unparsed cube-concept. The relation between cube-concepts allows to implement the intelligent sup-port of the user query forming based on joined data. At the stage "Process the integral OLAP-model" the integral OLAP-model defines set of dimensions and set of measures that can be analyzed together according to consistency requirement of the initial data sources and human cognitive ability. The considered stages are described in details below.

III. CONSTRUCTING THE INTERRELATED MULTIDIMENSIONAL MODEL OF THE NEW SOURCE

The main issue of the stage is an overcoming of heterogeneous of the joining data sources by constructing the data model for each source according to unified notation. The multidimensional model is an essential fundament for development of exploratory OLAP technology. Nowadays there are a lot of various solutions for discovering dimensions and measures in structured and semi-structured data. A survey of research achievements in the field of creating the multidimensional model for relational sources can be found in [5]. At the same paper the novel method of multidimensional design based on end-user requirements analysis is proposed. According to Winter et al. [6] the theoretical approaches in this field can be classified in supply-driven [7, 12], demand-driven [8] and hybrid methods [9]. The semi-automated methodology for multidimensional design of XML sources is proposed in [10]. In such areas as Social Network Analysis content-driven discovery of measurable facts and dimensional characteristics is based on applying data mining and other techniques to semi-structured data [11]. Great variety of researches in this field and its successful tests argue that the multidimensional data model is a good foundation for constructing comprehensive OLAP-model to join a set of heterogeneous sources. The similar problems are solving by data warehouse technology, but in this case the new data source assembling leans on ETL procedures which are developed manually. Proposed approaches and existing solutions give us the methodology base to develop the exploratory OLAP technology.

The specific feature of the proposing approach is to store information about the physical location and comparability of the analytical concepts for further automatic access to data and native model manipulation. In case of relational data sources the key demand is to keep original functional dependencies. The input of the stage «Constructing the interrelated multidimensional model of the new source» is the structure of the integrating data source as a metadata. The primary and foreign keys are considered for relational data sources. For data in XML format the data structure can be discovered by analyzing XSD files.

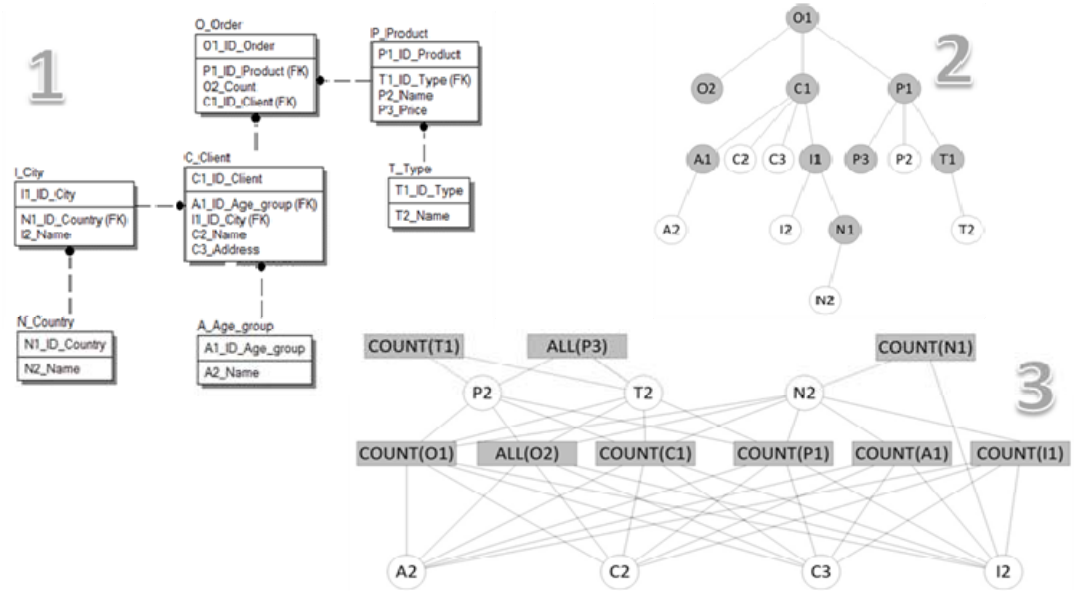

FIGURE II. CONSTRUCTING THE INTERRELATED MULTIDIMENSIONAL MODEL OF THE RELATIONAL SOURCE

This stage allows to differentiate dimensions from measures, to define possible aggregation functions for measures and to identify analytical relations between them. The output of the stage is the interrelated multidimensional model of the integrating source as a triplet (sD, sF, sR).

The Figure II illustrates the transformation of the relational schema (1) into graph of functional dependencies (2) and then into the interrelated multidimensional model (3) according to proposed algorithm [in print]. In figure the interrelated multidimensional model is presented in form of graph where the rectangle nodes correspond to measures sF with possible aggregation functions, the circle nodes correspond to dimensions sD and the edges show the available analytical relations sR.

IV. Integration of the Interrelated Multidimensional Model of the Source into the Common Model

The main idea of this stage is matching the interrelated multidimensional model of the joining source and the common interrelated multidimensional model. Applying the unified multidimensional notation for models matching gives us the significant advantage – comparing the dimensions is enough for analytical join of the heterogeneous resources. The dimensions can be considered as some connectors between information fragments in different formats and from different fields of the human activity. The input of this stage is the interrelated multidimensional model of the joining source, the output is the common interrelated multidimensional model as a triplet (gD, gF, gR). Initially, the common interrelated multidimensional model is the interrelated multidimensional model of the first source. This case is considering in the example.

Integration of the new source into the common model is based on hybridization of the modern approaches to logical and semantic matching of the database schemas (e.g. comparing the object trees, semantic and syntactic analysis of records) [13, 14]. The goal of this stage is development of the automatic matching method nevertheless it requires the human moderation obligatory. This process will be studded in the future work.

The proposed algorithm is intended to prove the authority approach to discovering the ordered multidimensional model and the way of constructing graph of MDM concepts. A normalized relational data base is subject of current research. To illustrate the algorithm execution we consider an example. Figure II represents studying relational data base schema. There are no partial dependencies on complex key and transitive dependencies in the normalized source.

V. Formation of the Matrix Context Based on the Common Interrelated Multidimensional Model

The integral analytical model of domain is based on Formal Concept Analysis (FCA) [15, 16] of the interrelated multidimensional model which can be presented as a binary matrix where the measures are the rows, the dimensions are the columns and relation of comparability is a cross at intersection between a row and a column.

The algorithm of the context formation is consequential considering the graph nodes of the common interrelated multidimensional model and constructing the reduced adjacency matrix. Figure III illustrates the context of the interrelated multidimensional model for studded example.

		C2	C3	P2	A2	I2	T2	N2
F1	COUNT(O1)	X	X	X	X	X	X	X
F2	ALL(O2)	X	X	X	X	X	X	X
F3	COUNT(C1)	X	X	X	X	X	X	X
F4	COUNT(P1)	X	X	X	X	X	X	X
F5	COUNT(A1)	X	X		X	X		X
F6	COUNT(I1)	X	X		X	X		X
F7	ALL(P3)			X			X	
F8	COUNT(T1)			X			X	
F9	COUNT(N1)					X		X

FIGURE III. THE CONTEXT OF THE INTERRELATED MULTIDIMENSIONAL MODEL

In Figure III the measures are named as gF = {F1, F2, F3, F4, F5, F6, F7, F8, F9}. The output of the stage is the context K of the common interrelated multidimensional model.

VI. Generation of the Integral Olap-Model

The stage “Generate the integral OLAP-model” is based on the author’s method of the integral OLAP-modeling and is aimed at defining the clusters that are presented as the unsparse analytical cube-concepts. Each cube-concept consists of measures with common dimensions and dimensions that characterize this set of measures.

The cube-concepts are formed based on the formal context K. For set A⊆F and for set B⊆D it is defined that: A’={d∈D | fRd for all f∈A} (all dimensions in D shared by the measures of A); B’={f∈F | fRd for all d∈B} (all measures in F can be processed with dimensions of B). The cube-concept is defined by derivation operators as pair (A, B) with A⊆F, B⊆D, A=B’, B=A’. It means that A is a set of equidimensional measures, which are processed with all dimensions of B. The set of measures A forms the extent and the set of dimensions B forms the intent of the cube-concept (A, B). The cube-concept is an analytical multidimensional cube which is complete with respect to addition of the equidimensional measures and the compatible dimensions.

The set of all cube-concepts is ordered by the subcube-supercube relation. For two cube-concepts (A1, B1) and (A2, B2) this order is identified as: (A1, B1)⩽(A2, B2) where A1⊆A2 and B2⊆B1. It means that the extent of a parent cube includes the extent of a child cube and the intent of the child cube includes the intent of the parent cube. In this case, (A1, B1) is called a subcube and (A2, B2) is called a supercube. The set of all concepts together with the subcube-supercube

relations forms a lattice L of the multidimensional cubes – the output of the stage.

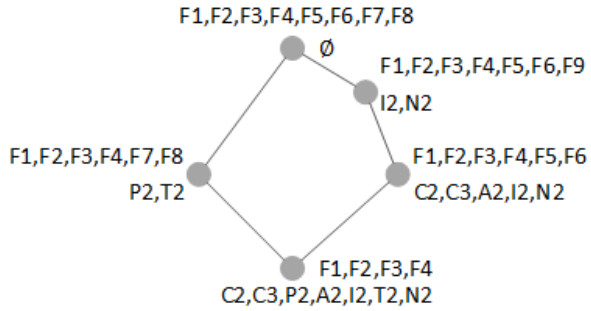

FIGURE IV. INTEGRAL OLAP-MODEL AS A LATTICE OF THE CUBE-CONCEPTS

Figure IV illustrates the lattice of the cube-concepts for studded example.

VII. PROCESSING OF THE INTEGRAL OLAP-MODEL

The lattice of the cube-concepts is an application of the integral OLAP-model of the domain. The integral OLAP-model covers all possible analytical queries of the do-main. The features of cube-concept lattice provide manipulation of all analytical concepts intuitively and support of the analytical exploring. During the analytical exploring the integral OLAP-model processing allows the user to see the set of analytical concepts (i.e. the set of dimensions and measures) that can be added to the query basing on the metadata of the original sources. When the user selects measures and (or) dimensions for analysis, the integral OLAP-model determines the set of cube-concepts that correspond to user's query and form a sub-lattice. The cube-concept of sub-lattice located on the upper level (supreme) contains the maximal set of the addi-tional measures and the cube-concept located on the lowest level (infimum) contains the maximal set of the additional dimensions. Changing the query leads to adaptation of the sub-lattice according to original algorithm.

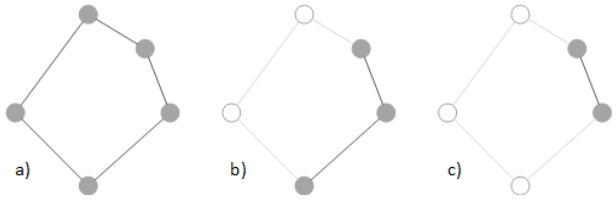

FIGURE V. PROCESS OF THE SUB-LATTICE ADAPTATION

Figure V illustrates the process of the sub-lattice adaptation to the user analytical interests. At the first step, the user query is empty and the sets of the additional analytical concepts are equal to entire set F and D respectively. The user analytical query corresponds to the lattice of the integral OLAP-model (Figure V.a). In our example the second step query is (Ø,{N2}). It can be executed by 3 cube-concepts (Figure V.b). At the next step the measure F6 has been added to the query and the sub-lattice has been updated. The current query corresponds to 2 cube-concepts that give the user the sets of additional measures F1, F2, F3, F4, F5, F9 and dimensions C2, C3, A2, I2 (Figure V.c).

VIII. CONCLUSION

The paper presents the original approach to the exploratory OLAP of the heterogene-ous resources based on the integral analytical model. Applying the multidimensional data model as a technological fundament allows to analyze the structure of the joining sources according to unified notation, allows to present information in form of native multidimensional view and provides efficient integration of the sources by matching the dimensions. Storing information about the physical location and comparability of the analytical concepts within multidimensional model enables to form automatically the exploratory queries. The proposed approach leans on analysis and matching resources on the conceptual level (i.e. without physical data consolidation) that allow to process data with a lot of records and entities. In the paper, the requirements to exploratory OLAP technology are defined and the main issues and their solutions are formulated based on preliminary scientific results. The principal stages for analytical exploration support according to author's method of the integral OLAP-modeling are described.

According to the proposed approach the future work is related with the following tasks: development of the algorithms for the interrelated multidimensional model forming in the case of XML, JASON, RDF data formats; development of the methods and algorithms for the interrelated multidimensional models matching and development of the interface for the integral OLAP-model manipulation.

ACKNOWLEDGMENT

The reported study was funded by RFBR according to the research project №16-07-01001 and by RFBR and Government of Krasnoyarsk Territory according to the research projects №16-41-240425.

REFERENCES

[1] E.F. Codd, S.B. Codd, C.T. Salley Providing OLAP. On-line Analytical Processing to Us-er-Analists: An IT Mandate, Codd & Associates; 1993.

[2] L.F. Nozhenkova, V.V. Shidurov "OLAP-technologies of operational analytical support of administration", Information technologies and computer systems, 2010, No 2, pp.15-27.

[3] A. Abello, O. Romero, T.B. Pedersen, R. Berlanga, V. Nebot, M.J. Aramburu, A. Simitsis, "Using semantic web technologies for exploratory OLAP: a survey", IEEE Transactions on Knowledge and Data Engineering, 27(2), 571-588.

[4] T. Penkova, A. Korobko, "Method of constructing the integral OLAP-model based on formal concept analysis", Frontiers in Artificial Intelligence and Applications, Vol. 243, 2012, pp. 219-227, doi:10.3233/978-1-61499-105-2-21.

[5] O. Romero and A. Abelló, "Automatic validation of requirements to support multidimen-sional design," Data & Knowledge Engineering, vol. 69(9), 2010, pp. 917-942.

[6] R. Winter, B. Strauch, "A method for demand-driven information requirements analysis in data warehousing projects," Proceedings of the 36th Annual Hawaii International Confer-ence on System Sciences, IEEE, 2003, pp. 231-239.

[7] M. Golfarelli, D. Maio, and S Rizzi, “The dimensional fact model: A conceptual model for data warehouses,” International Journal of Cooperative Information Systems, 7.02n03, 1998, pp. 215-247.

[8] P. Giorgini, S. Rizzi, M. Garzetti, Goal-oriented requirement analysis for data warehouse design, Proc. of 8th Int. Workshop on Data Warehousing and OLAP, ACM Press, 2005, pp. 47–56.

[9] C. Phipps, K.C. Davis, “Automating data warehouse conceptual schema design and evalu-ation,” Proc. of 4th Int. Workshop on Design and Management of Data Warehouses, vol. 58, 2002, pp. 23 –32.

[10] B. Vrdoljak, M. Banek, and S. Rizzi, “Designing Web Warehouses from XML Schemas,” Proc. Fifth Int’l Conf. Data Warehousing and Knowledge Discovery (DaWaK ’01), 2003, pp. 89-98.

[11] S. Mansmann, N.U. Rehman, A. Weiler and M.H. Scholl, “Discovering OLAP dimensions in semi-structured data,” Information Systems, vol. 44, 2014, pp. 120-133.

[12] M.R. Jensen, T. Holmgren, T.B. Pedersen, “Discovering multidimensional structure in re-lational data,” 6th Int. Conf. on Data Warehousing and Knowledge Discovery, LNCS, vol. 3181, Springer, 2004, pp. 138–148.

[13] S. Melnik, A. Adya, P.A. Bernstein, “Compiling mappings to bridge applications and data-bases”, ACM Transactions on Database Systems (TODS), 2008, T. 33, №. 4, C. 22.

[14] H.H. Do, E. Rahm, “Matching large schemas: Approaches and evaluation”, Information Systems, 2007, T. 32, №. 6, C. 857-885.

[15] R. Wille, Restructuring Lattice Theory: an approach based on hierarchies of concept, Reidel, Dordrecht-Boston, 1982.

[16] B. Ganter, R. Wille, Formal Concept Analysis: mathematical Foundations. Springer-Verlag, Berlin Heidelberg New York, 1999.