UDC 681.3.016

Objects of Interest Detection by Earth Remote Sensing Data Analysis

Anastasia V. Demidova[*], **Maxim B. Fomin**[†], **Sergey G. Shorokhov** [†]

[*] *Department of Applied Probability and Informatics,*
Peoples' Friendship University of Russia (RUDN University),
6 Miklukho-Maklaya str., Moscow, 117198, Russian Federation
[†] *Department of Information Technologies*
Peoples' Friendship University of Russia (RUDN University)
6 Miklukho-Maklaya str., Moscow, 117198, Russian Federation

Email: demidova-av@rudn.ru, fomin_mb@rudn.university, shorokhov_sg@rudn.university

In information systems based on a multidimensional approach, a data model is a multidimensional data cube. If one uses a large set of aspects for the analysis of data domain the data cubes are characterized by substantial sparseness. This makes it difficult to describe the metadata of the information system and complicates the organization of data storage. To describe the structure of a sparse data cube, a cluster method can be used. This method is based on the construction of groups of members which are semantically connected with other groups of members. Connected groups related to different dimensions describe the cluster of cells. Classification schemes that correspond to the structural components of the observed phenomenon can be used to describe it's semantics. Every classification scheme is a graph describing the hierarchy of members that are associated with a separate structural component of the observed phenomenon. The coupling between several classification schemes related to different structural components helps to describe the metadata of the multidimensional information system. Classification schemes are a source of classification of information objects of a multidimensional cube related to the structural components of the observed phenomenon.

Key words and phrases: OLAP, data warehouse, multidimensional data model, sparse data cube, set of possible member combinations, cluster of member combinations.

In: K. E. Samouylov, L. A. Sevastianov, D. S. Kulyabov (eds.): Selected Papers of the 1st Workshop (Summer Session) in the framework of the Conference "Information and Telecommunication Technologies and Mathematical Modeling of High-Tech Systems", Tampere, Finland, 20–23 August, 2018, published at http://ceur-ws.org

1. Introduction

Multidimensional information systems based on the principles of OLAP are used for the operational analysis of large datasets. Analytical space in a system of this type is a multidimensional data cube. The role of the cube dimensionalities is played by the dimensions corresponding to various aspects of the observed phenomenon for which description the system is developed. If we use a large amount of semantically heterogeneous data for the description of the observed phenomenon the multidimensional cube is characterized by high sparseness and irregular filling [1–8]. As a result, there is a problem of developing an adequate way to describe the structure of an analytical space which use would make it possible to effectively organize the data analysis process [9–17]. Such a correct way should provide the accounting of semantics of the observed phenomenon.

The cluster method can be used for the effective description of the multidimensional cube structure. This method is based on the semantic analysis of different dimensions' members' compatibility in possible cube cells. It allows describing the metadata of the information system as a set of possible member combinations. Possible combinations comply with possible cells of the multidimensional cube. Every possible cube cell complies with some fact.

Difficulties in describing the structure of the analytical space may arise in case if, in the process of forming the metadata of the information system, the analysis of the semantic aspects of the observed phenomenon is subject to technological aspects. The observed phenomenon is a set of interrelated processes related to the subject domain. Data describing the observed phenomenon can form one or more multidimensional data cubes. In describing the structure of an analytic space, the following problems must be solved:

- the problem of classification of data describing the observed phenomenon;
- the problem of accounting for the semantics of the observed phenomenon in these data.

2. Multidimensional data model

The structure of multidimensional data model should reflect the aspects of subject domain which are used in the data analysis process. Each aspect corresponds to one dimension of a multidimensional cube H. A full set of dimensions forms a set $D(H) = \left\{D^1, D^2, .., D^n\right\}$, there D^i is i-dimension, and $n = \dim(H)$ – dimensionality of multidimensional cube [18]. Each dimension is characterized by a set of members $D^i = \left\{d_1^i, d_2^i, .., d_{k_i}^i\right\}$, there i is a number of dimension, k_i – the quantity of members. Members of D^i are drawn from a set of positions of the basic classifier which corresponds to an aspect of the observed phenomenon associated with D^i.

The multidimensional data cube is a structured set of cells. Each cell c is defined by a combination of members $c = \left(d_{i_1}^1, d_{i_2}^2, .., d_{i_n}^n\right)$. The combination includes one member for each of the dimensions. If the analysis of the observed phenomenon is performed using a large set of diverse aspects, not all members combinations define the possible cells of multidimensional cube, i.e. the cells corresponding to a certain fact. This effect occurs due to semantic inconsistencies of some members from different dimensions to each other and generates sparseness in the cube.

The complex structure of the compatibility of members may lead to a situation where a certain dimension becomes semantically uncertain if combined with a set of members from other dimensions. In this situation, while describing the possible cell of multidimensional cube the special value "Not in use" can be used to set the member of semantically unspecified dimension. The structure of the multidimensional data cube in the information system can be described as the set of possible members combinations. Different values from the classifiers, which comply with the dimensions, and the special

value "Not in use" can be applied in the combinations of this set. To refer to the set of possible members combinations we will use the abbreviation "SPMC".

The subject domain is characterized by the measure values defined in possible cells of the multidimensional cube. The full set of measures composes the set $V(H) = \{v_1, v_2, .., v_p\}$, where v_j is j-measure, p – the quantity of measures in the hypercube. Not all the measures from the $V(H)$ can be defined in the possible cell. This situation can appear in case of semantic inconsistency between the members defining the cell and some measures. While describing multidimensional data cube structure for every possible sell it is necessary to define its own set $V(c) = \{v_1, v, .., v_{p_c}\}$, which consists of certain measures for this cell, $1 \leq p_c \leq p$. We can use the special value "Not in use" for the description of c measures, which are not included in the set $V(c)$.

The description of the SPMC can be obtained with the help of the cluster method based on the analysis of links between members [19]. The cluster method allows identifying the groups of members. The group $G_j^i = \left\{d_1^i, d_2^i, .., d_{m_j}^i\right\}$ of members in i-dimension includes m_j members ($1 \leq m_j \leq k_i$), where j is a group number and contains members, which equally coincide in the SPMC with the members from some groups of members of other dimensions.

It is possible to define connected groups of member in different dimensions with the help of the semantic analysis. The cluster of members combinations K is the set of member combinations, which can be obtained with the help of Cartesian product where operands are groups of members or special value "Not in use"; one operand stands for every dimension used in the cluster $\text{SPMC}(K) = G_1 \times G_2 \times .. \times G_n$. Clusters of members combinations can be used for the description of the SPMC.

3. The use of classification schemes for describing the semantics of the observed phenomenon

From the position of semantic the description of the observed pattern characteristics within the multidimensional data model consists of the classification attributes detecting (dimensions of the multidimensional cube) and establishing the links between them. It can be rather difficult if there are a great number of dimensions. Classification schemes (SC) can be used to solve the problem of classification of data describing the observed phenomenon. For CS it is possible to formulate a number of requirements [20].

It is necessary to take into account the component structure of the observed pattern while defining CS. If the observed pattern can be semantically divided into separate structural components for which is possible to choose their own sets of aspects for analysis every component should be compared with CS. The procedure of CS formation is based on defining and analysis of the attributes relevant to the chosen aspects of the analysis. The dimensions of the multidimensional cube should be compared with the characteristics.

CS of the attributes for the observed patterns should be formed on the hierarchical principle. Ranking should be established among the attributes related to CS. This ranking allocates the dimensions which to some extent convey the essence of the structural component for the observed pattern. This component is compared with CS. It is necessary to define the major dimension which is more likely to reflect the semantic of the structural component relevant to CS. The hierarchy of attributes should be formed from other dimensions included into CS which are semantically subordinate to the major dimension and express some particular properties of the structural component for the observed pattern. The following principle should be observed: the members of the major dimension convey the most important attributes of the observed pattern, the members of other dimensions which come hierarchically below the major one convey some subordinate attributes specifying the essence of the major dimension.

While forming the hierarchy of the attributes for the observed pattern in CS it should be possible to describe the members of the major dimension separately or in groups of members as different members can be connected with different semantic aspects of the

structural component for the observed pattern. Different hierarchies of attributes should be formed for the members of the major dimension which are semantically different.

In the hierarchy of the attributes in CS there must be the information about the set of measures describing the observed pattern in case of choosing some particular members from the hierarchy.

The classification scheme of the attributes for the observed pattern is an object of the multidimensional information system which describes the structural component of the observed pattern and contains the following information:

- the set of dimensions included into the classification scheme;
- the set of members of these dimensions included into the classification scheme;
- the major dimension chosen in the set of dimensions CS;
- the set of measures included into the classification scheme;
- the tree of member combinations CS which form the hierarchy of the attributes included into CS.

The hierarchy principal of CS forming is realized in the structure of a tree which presents the member combinations in CS. The tree of combinations can be formed as a result of the semantic analysis of the structural component for the observed pattern. The tree can be defined while describing the process of its formation. One should start the formation of the tree from its roots where the groups of members of the major dimension are placed. Then it is necessary to go down passing the hierarchical levels and adding a group of members to each of them. Thus every group reveals the essence of every previous member on the previous level. What is more it is necessary to add the group related to the dimension which is mostly connected with the members of the previous level. As a result different sequences of CS measures can appear in tree brunches on the way from the roots to leaves.

Relationships between the members of different dimensions can be established using approaches of non-parametric methods of statistical analysis [21] and queueing theory [22].

For the tree structure of classification scheme, the following rules must be followed:

1. The root of the tree is the unit “Major dimension”.
2. The tree itself is a hierarchical structure where the levels are set through alternating such units as “Group of members” and units “Dimension”. At the same time groups of members should be formed in the dimensions relevant to the units hierarchically placed one level higher.
3. Leaves of the tree are units “Group of members”.
4. The unit “Group of members” (except the unit which is a tree leaf) should be relevant to the unit “Dimension” hierarchically placed on a lower level. Only one unit or several units “Group of members” placed on a lower level can be relevant to the unit “Dimension”.
5. Moving from the root to a leaf you can see every dimension only once.

4. Semantic aspects of information system metadata construction

The analysis of the observed phenomenon reveals the qualification characteristics that are included in the metadata of the information system. In the structure of a multidimensional data cube, these characteristics are divided into subject of analysis (measures) and aspect of analysis (dimensions). Semantic analysis allows to establish connections between these characteristics. As a result, the structure of a multidimensional cube can be revealed, that is, significant cells of a multidimensional cube corresponding to the facts are described.

Difficulties in the application of the described technique are due to the fact that the complex observed phenomenon is characterized by a large number of aspects. A complete set of these aspects allows us to construct many multidimensional data cubes corresponding to different structural components of the observed phenomenon. The pairwise analysis of the characteristics does not make it possible to separate them in accordance with the structural components, since they are mixed in the observed phenomenon and there are no hierarchical relationships between them.

Construction of classification schemes related to the observed phenomenon as a result of semantic analysis allows to achieve the following result:

- semantic separation of the characteristics of the observed phenomenon, their binding to the structural components of the observed phenomenon;
- ranking of characteristics, building a hierarchy of characteristics in accordance with their significance in the description of the properties of the structural component of the observed phenomenon.

The described properties of classification schemes allow us to consider them as the main objects that describe and systematize information about the structural components of the observed phenomena. The interaction of information objects included in the multidimensional data cube and the classification scheme can be represented by a diagram (see figure 1).

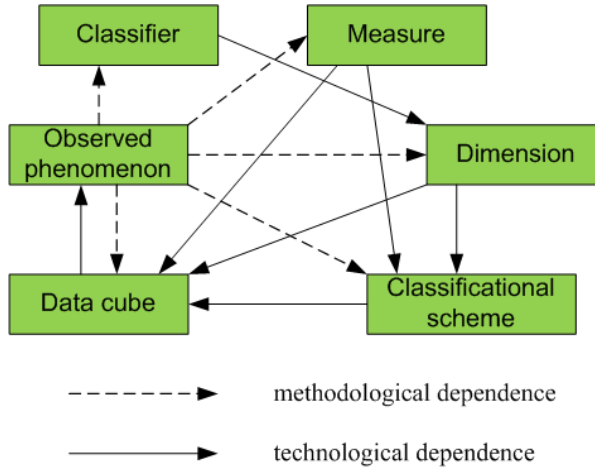

Figure 1. Information objects describing the observed phenomenon

The links in the diagram represent semantic and technological dependencies. The use of classification schemes alters the relationship between the observed phenomenon and the multidimensional data cube. Classification schemes are a source of classification of information objects of a multidimensional cube related to the structural components of the observed phenomenon.

Members belonging to different dimensions and measures form a classification scheme. The process of such formation is influenced by the semantic relationship between the observed phenomenon on the one hand and the dimensions and measures on the other hand. Due to the fact that the classification scheme expresses the properties of the structural component of the observed phenomenon, there is an implicit semantic relationship between the classification scheme and the observed phenomenon, which forms the composition of dimensions, members and measures of the classification scheme. The same relationship allows us to determine which classification schemes can underlie multidimensional data cubes related to the observed phenomenon. The insertion of classification schemes changes the nature of the relationship between the observed phenomenon and the multidimensional data cube. The observed phenomenon loses its role as a source of classification, while the multidimensional cube plays the role of a technological component.

5. Conclusions

The paper considers the method of designing information systems using a multidimensional approach. This approach allows us to develop a system based on a metamodel, which is semantically related to the subject domain of the system. The method is based on the construction of groups of members which are semantically connected with other groups of members. Connected groups related to different dimensions describe the cluster of cells.

Classification schemes that correspond to the structural components of the observed phenomenon can be used to describe it's semantics. Every classification scheme is a graph describing the hierarchy of members that are associated with a separate structural component of the observed phenomenon. Classification schemes can be used to solve the problem of classification of data describing the observed phenomenon. The coupling between several classification schemes related to different structural components helps to describe the metadata of the multidimensional information system. It is formed on the hierarchical principle and establishes a ranking between the characteristics of structural component of observed phenomenon. The classification scheme is a technological component in relation to multidimensional data cube, while the multidimensional cube plays the technological role in relation to the observed phenomenon.

Acknowledgments

The publication has been prepared with the support of the "RUDN University Program 5–100".

References

1. E. Thomsen, OLAP Solution: Building Multidimensional Information System, Willey Publishing, 2002.
2. C. M. Hirata, J. C. Lima, Multidimensional cyclic graph approach: representing a data cube without common sub-graphs. Inf. Sci. 181, pp. 2626–2655 (2011).
3. N. Karayannidis, T. Sellis, Y. Kouvara, CUBE File: A File Structure for Hierarchically Clustered OLAP Cube, in: Advances in Database Technology, pp. 621–638, Springer-Verlag (2004) ISBN 978-3-540-21200-3.
4. S. Chun, Partial Prefix Sum Method for Large Data Warehouses, in: Foundations of Intelligent Systems — ISMIS 2003, pp. 473–477. Springer-Verlag (2004) ISBN 978-3-540-39592-8.
5. R. B. Messaoud, O. Boussaid, S. L. Rabaseda, A Multiple Correspondence Analysis to Organize Data Cube, in: Databases and Information Systems IV — DB&IS 2006, pp. 133–146. IOS Press (2007) ISBN 978-1-58603-715-4.
6. R. Jin, J. K. Vaidyanathan, G. Yang, G. Agrawal, Communication and memory optimal parallel data cube construction, IEEE Transactions on Parallel and Distributed Systems. 16, pp. 1105–1119 (2005).
7. Z. W. Luo, T. W. Ling, C. H. Ang, S. Y. Lee, B. Cui, Range Top/Bottom k Queries in OLAP Sparse Data Cubes, in: Database and Expert Systems Applications — DEXA'01, pp. 678–687. Springer-Verlag (2001) ISBN 978-3-540-42527-4.
8. L. Fu, Efficient Evaluation of Sparse Data Cubes, in: Advances in Web-Age Information Management — WAIM'04, pp. 336–345. Springer-Verlag (2004) ISBN 978-3-540-27772-9.
9. F. Z. Salmam, M. Fakir, R. Errattahi, Prediction in OLAP Data Cubes. Journal of Information & Knowledge Management, 15, pp. 449–458 (2016).
10. 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 & Data Engineering, 27, pp. 571–588 (2015).
11. L. I. Gomez, S. A. Gomez, A. Vaisman, A generic data model and query language for spatiotemporal OLAP cube analysis, in: Proceedings of the 15-th International

Conference on Extending Database Technology — EDBT 2012, pp. 300–311. ACM, New York (2012) ISBN 978-1-4503-0790-1.
12. M. F. Tsai, W. Chu, A Multidimensional Aggregation Object (MAO) Framework for Computing Distributive Aggregations, in: Data Warehousing and Knowledge Discovery — DaWaK 2003, pp. 45–54. Springer-Verlag (2003) ISBN 978-3-540-40807-9.
13. J. S. Vitter, M. Wang, Approximate computation of multidimensional aggregates of sparse data using wavelets, in: Proceedings of the 1999 International Conference on Management of Data — SIGMOD'99, pp. 193–204. ACM, New York (1999), ISBN 1-58113-084-8.
14. B. Leonhardi, B. Mitschang, R. Pulido, C. Sieb, M. Wurst, Augmenting OLAP Exploration with Dynamic Advanced Analytics, in: Proceedings of the 13th International Conference on Extending Database Technology — EDBT 2010, pp. 687–692. ACM (2010) ISBN 978-1-60558-945-9.
15. W. Wang, H. Lu, J. Feng, J. X. Yu, Condensed Cube: An Effective approach to reducing data cube size, in: Proceedings of the 18th International Conference on Data Engineering — ICDE'02, pp. 155–165. IEEE Computer Society (2002) ISBN 0-7695-1531-2.
16. S. Goil, A. Choudhary, Design and implementation of a scalable parallel system for multidimensional analysis and OLAP, in: Parallel and Distributed Processing — 11th IPPS/SPDP'99, pp. 576–581. Springer-Verlag (1999) ISBN 978-3-540-65831-3.
17. A. Cuzzocrea, OLAP Data Cube Compression Techniques: A Ten-Year-Long History, in: Future Generation Information Technology — FGIT 2010, pp. 751–754. Springer-Verlag (2010) ISBN 978-3-642-17568-8.
18. C. Chen, J. Feng, L. Xing, Computation of Sparse Data Cubes with Constraints, in: Data Warehouse and Knowledge Discovery, pp. 14–23. Springer-Verlag (2003) ISBN 978-3-540-40807-9.
19. M. B. Fomin, Cluster method of description of information system data model based on multidimensional approach. In: Vishnevskiy V., Samouylov K., Kozyrev D. (eds.) Distributed Computer and Communication Networks. DCCN 2016. CCIS, vol 678, pp. 657–668, Springer (2016). doi:10.1007/978-3-319-51917-3_56.
20. M. B. Fomin, The application of classification schemes while describing metadata of the multidimensional information system based on the cluster method. In: Vishnevskiy V., Kozyrev D. (eds.) Distributed Computer and Communication Networks. DCCN 2017. CCIS, vol 700, pp. 307–318, Springer (2017). doi:10.1007/978-3-319-66836-9_26.
21. Y. Orlov, Y. Gaidamaka, E. Zaripova, Approach to estimation of performance measures for SIP server model with batch arrivals, in: Communications in Computer and Information Science, vol 601 (2016) pp. 141–150. doi:10.1007/978-3-319-30843-2_15.
22. Y. V. Gaidamaka, E. R. Zaripova, Session setup delay estimation methods for IMS-based IPTV services. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol 8638, LNCS, 2014, pp. 408–418. doi:10.1007/978-3-319-10353-2_36.