Exploring the Properties of the Context and Lattice of the Integral Analytical Model*

Anna Korobko[0000−0001−5337−3247]

Institute of Computational Modeling of the Siberian Branch
of the Russian Academy of Sciences, 50/44 Akademgorodok, Krasnoyarsk, 660036, Russia
lynx@icm.krasn.ru

Abstract. Diversity, multidimensionality, and the amount of available information require an original approach to the operational analytical processing of heterogeneous data. The integral analytical model provides a federation of heterogeneous data without physical moving, with the support of interactive visual exploration of a large model, and with the execution of analytical queries on distributed data sources simultaneously. Compact representation and native management of big data is achieved by presenting the model in the form of a context and building a lattice for it, in accordance with the FCA method. The theory of integral analytical modeling (IAM) relies on the fact that the context of the model has special properties that ensure fast construction of the lattice and its compactness. The goal of the article is to conduct a comparative analysis of the properties of the IAM context and contexts of various origins, to evaluate and compare the rate of the lattice generation and their properties.

Keywords: Context Properties, OLAP, FCA, Exploratory OLAP, Heterogeneous Data, Big Data.

1 Introduction

The value of open information resources provides a unique opportunity to make effective management decisions based on a broader factual base [1]. In response to demand, the data analysis software market is actively developing. The capabilities of domestic (Yandex.DataLens, Polymatica, Visiology и др.) and foreign (Tableau, Qlik, MS Power BI, etc.) data analysis systems vary from methods of mathematical statistics to analytical platforms with built-in methods of data mining. Choosing a data analysis system, companies first pay attention to price and functionality. Secondly, they look for design clarity and a user-friendly interface, and rapid data access. The speed of including new data into the analysis process and reduction in user requirements are becoming important options.

Information technologies are dynamically developing and provide new forms of data presentation and methods of their processing. The wide choice of analytical systems and the constant development of new software confirm this. Diversity, multidimensionality, and the amount of available information require the development of original approaches to the operational analytical processing of heterogeneous data. The world scientific community formulates various aspects of the problem as separate tasks: Exploratory On-Line Analytical Processing (ExOLAP) [2], Self-service Business Intelligent (Self-service BI) [3] and Big Data [4].

The integral analytical model (IAM) provides a federation of heterogeneous data without physically moving, with the interactive exploration of a large model, and with the execution of analytical queries on distributed multiple data sources simultaneously. The theory of building IAM is based on the technology of online analytical data processing (OLAP) and method of formal conceptual analysis (FCA). The main requirement of the OLAP technology is the presentation of data in a multidimensional view. Categorical data with a finite domain of values are related to dimensions, and numerical data are called measures. The multidimensional representation has an impressive theoretical base. It is widely used in modern data analysis systems. A lot of popular tools for visualizing the analysis results are built on its basis. The IAM construction method consists in virtual combining (integration) of heterogeneous data based on the theory of multidimensional modeling. The structure of the combined sources is described in terms of a multidimensional representation and integrated into a general integral model. The FCA method has become an elegant solution of the problem of research and management of a wide integral analytical model. As a result of the adaptation of the method to the terms of multidimensional data representation, it is possible to present the integral analytical model in the form of an algebraic lattice with OLAP cubes at the vertices.

The proposed approach has improved significantly over the last 10 years. Original methods have been developed for constructing multidimensional models for relational sources and databases of XML documents [5], a method for combining models of heterogeneous sources has been proposed [6], a method for constructing an IAM in the form of a lattice of OLAP cubes has formally been described, a method has been proposed to support the formation of a user analytical query to integral models and models for a number of subject areas have been built (prevention and elimination of emergency consequences, effectiveness of scientific activity and support for placing a municipal procurement). However, the answer to the key question has been outside the scope of the study. Does the representation of the integral model as a lattice satisfy the requirement for the efficiency of analytical processing of a combined set of heterogeneous data?

2 Problem Statement

The result of the analytical integration of heterogeneous sources is a binary matrix of bi-adjacency. Matrix rows correspond to measures, columns to dimensions, and the cell value at the intersection of a row and a column indicates the analytical

compatibility of the elements. By interpreting the constructed matrix as a context, we can construct a lattice for it. This paper is devoted to the study of the IAM properties of the speed of building a model by the context and properties of the model as a lattice of OLAP-cubes. The study was conducted in the form of a computational experiment. The object of research is IAM for placing a municipal procurement. The model is built using the analytical integration software module and combines two dissimilar sources: relational (Regional system of forming demands) and XML-documents base (Unified information system in the field of procurement). The model context ("IAM_context") contains 1442 rows and 263 columns. The percentage of matrix filling is 2.65% - 10,046 non-zero values.

The main scientific hypothesis is that the IAM context has special properties that ensure the rapid construction of the lattice and its compactness. It means that IAM is suitable for supporting the on-line analytical processing of big heterogeneous data and rapid integration of new sources. The aim of the article is to perform a comparative analysis of the parameters of contexts of different origin, to evaluate and compare the rate of the lattice generation and size for different contexts.

Context and algebraic lattice construction are the core of many modern decision support methods [7]: information retrieval, classification, formation of recommendations, generation of association rules etc. Researchers use the FCA method to study the structural features of texts, user preferences, ontological concepts, and purchases. A classic example of the capabilities of association rules is the market basket analysis. Consider the Instacart Market Basket Analysis dataset as an example of a real domain for the context construction and comparison with the IAM context. The dataset contains information on orders in the Instacart grocery delivery service. The data was downloaded from the public platform of the data analysis competition – Kaggle.com (https://www.kaggle.com/psparks/instacart-market-basket-analysis). The dataset consists of six files; to form a binary context, we need only one - order_products__prior.csv. The file describes the correspondence of the order identifiers and product codes. For the experiment, the data is loaded as a Pandas DataFrame and converted to the binary context ("IMBA_context").

The context rows correspond to orders and the columns correspond to purchased products. The resulting context is larger than the "IAM context". We limited the original dataset and considered two contexts: ("IMBA_context_1") coinciding in the number of non-zero elements with the IAM context and ("IMBA_context_2") close in size. The size of the context "IMBA_context_1" is 980 rows and 4521 columns, the percentage of filling is 0.23%. The size of the context "IMBA_context_2" is 238 rows and 1596 columns, the percentage of filling is 0.59%. These contexts differ in that the number of columns is greater than the number of rows. In the context of "IAM_context" we can see the opposite proportions. In term of the experimental integrity, we considered additionally transposed contexts – "IMBA_T_context_1" and "IMBA_T _context_2". The control context ("RND_context") is generated using a random number generator. "RND_context" has the same size and filling density as the "IAM_context".

3 Experimental Study

The computational experiment was implemented in the JupyterLab environment in Python 3.7 using the libraries: pandas, numpy, matplotlib, plotly, net-workx and time. To calculate the lattice concepts, we used the original implementation of the "In Close" [8] algorithm, optimized using the built-in data structures of the Python language. Link to the project page is https://github.com/khroom/FCA_LAB. The generation of concepts for all the contexts was performed by a single function with the measurement of the computation time (Fig.1).

```
def is_cannonical(self, column, new_a, r):
    for i in range(column, -1, -1):
        if self.context.columns[i] not in self.concepts[r]['B']:
            if new_a.issubset(self.context_derivation_1.iloc[i]):
                return False
    return True


def in_close(self, column: int, r: int, threshold=0.0):
    for j in range(column, len(self.context.columns)):
        new_concept = {'A': self.context_derivation_1.iloc[j].intersection(self.concepts[r]['A']),
                       'B': set()}
        if len(new_concept['A']) == len(self.concepts[r]['A']):
            self.concepts[r]['B'].add(self.context.columns[j])
        else:
            if (len(new_concept['A']) != 0) and
            (len(new_concept['A']) > self.threshold_base * threshold):
                if self.is_cannonical(j - 1, new_concept['A'], r):
                    new_concept['B'] = new_concept['B'].union(self.concepts[r]['B'])
                    new_concept['B'].add(self.context.columns[j])
                    self.concepts.append(new_concept)
                    self.in_close(j + 1, len(self.concepts) - 1, threshold)
```

Fig. 1. The python code of the “In-Close” algorithm.

A comparative analysis of the following aspects was conducted: properties of contexts, speed of the generation of a lattice by the context, number of concepts (vertices) of a lattice and properties of concepts (extent and intent). The comparison results are shown in Table 1.

Table 1. The comparison results.

	IAM_ context	IMBA_ context_1	IMBA_T_ context_1	IMBA_ context_2	IMBA_T_ context_2	RND_ context
Size	1,442x263	980x4521	4,521x980	238x1596	1,596x238	1,442x263
Fill density	2.65%	0.23%	0.23%	0.59%	0.59%	2.64%
Number of non-zero items	10,046	10,046	10,046	2,255	2,255	10,046
Speed of concepts generation	1.36s	9min 45s	2min	25.1s	2.74s	1min 2s
Number of concepts	205	6722	6722	642	642	10633
Maximal extent of concepts	394 (27.3%)	158 (16.1%)	46 (1%)	28 (11.8%)	34 (2.1%)	54 (3.7%)

Maximal intent of concepts	79 (30%)	46 (1%)	158 (16.1%)	34 (2.1%)	28 (11.8%)	15 (5.7%)

The research results show that the used algorithm is sensitive to the ratio of the context sizes. The generation time for the concepts for the transposed contexts “IMBA_T_context_1” and “IMBA_T_context_2” is significantly lower than that for the original contexts. This can be used to optimize the algorithm. Figure 2 shows a fragment of a scatter diagram of the extent and intent of concepts.

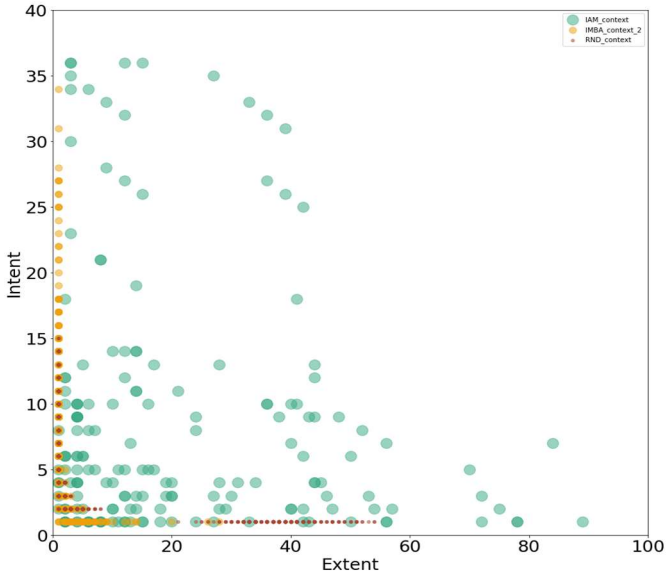

Fig. 2. The Scatter plot of the concept intents and extents.

We interpret the speed of the generation of concepts and their number as a sign of the presence or absence of a special structure in the context. The IAM context is superior to other contexts in both parameters. The most significant differences are between the IAM context and a randomly filled context. The context “RND_context” is filled evenly with a given density, which is reflected in the generation speed of 1 minute 2 seconds, and in the number of concepts of 10633. A large number of the lattice concepts imply an unstructured context.

The scatterplot shows that when the extent is 1, the intent changes from 21 to 54 for the “RND_context”. And when the intent is 1, the extent belongs to the interval (2, 15). The extent and intent of the rest of the concepts do not exceed 8. The artificial context is not a domain model and does not reveal relationships between the real entities. While the IAM context lattice is built in 1.36 seconds, it has 205 concepts and contains the concepts that unite up to 30% of the rows and columns. Fast generation, concepts large in extent and intent, and compactness of the lattice suggest a strong structuring of the context - the integral analytical model of heterogeneous sources.

Now we will consider the contexts built for another real domain - grocery delivery service. Let us consider transposed versions of the contexts, as this significantly affects the speed of the concept generation. The context “IMBA_T_context_1” and the context “IAM_context” have the same number of nonzero elements but differ in size. The “IMBA_T_context_1” is 11 times larger and 11 times sparser than the “IAM_context”. The concept generation time for the “IMBA_T_context_1” is 2 minutes, and the lattice includes 6722 concepts.

The context “IMBA_T_context_1” is semi-structured. The number of the concepts is lower than that of the “RND_context”, but due to the large size, the generation time is significantly longer. The significant difference between the maximum extent and the maximum intent of the concepts indicates a specific structure of the “IMBA_T_context_1”. We can see a widespread among the purchased products and a weak relationship between them. Due to the nature of the subject area, researchers often study product groups to solve the problem of defining related products. The context "IMBA_T_context_2” has a size comparable to the size of “IAM_context”. It is based on the “IMBA_T_context_1” but has a smaller size and higher density. The concept generation time for the “IMBA_T_context_2” is 2.74 seconds, and the lattice includes 642 concepts.

The context “IMBA_T_context_2” is also semi-structured. The number of the concepts is 3 times greater than the number of the concepts in the IAM context, with a similar dimension. The difference between the maximum extent and intent of the concepts is not as significant as in the larger concept. This means that the data is not homogeneous. Figure 2 shows that for this context, many concepts have an extent equal to 1 and larger sized intents - these are receipts that combine up to 34 products. There are not many popular products and there are even fewer repetitive product combinations. Faster generation of the concepts compared to a random context implies the structuredness of the context.

The results of the comparative analysis of the generation parameters and properties of the contexts confirm that the IAM context has a special structure. These special properties make the lattice fast and compact. Given the same size, number of nonzero elements and density of the context, the estimated parameters are greatly different and strongly affect the efficiency of manipulating the concept lattice. This means that the compactness and speed of constructing the lattice is determined by some internal properties of the context, structural relationships of the entities of the modeled subject area.

During the study, significant structural features of the real IAM context were identified. The functional dependences between the attributes of the original storage schemes are reflected in the hierarchical dependences between the dimensions of the multidimensional model. In the context of IAM, they take the form of “mutual existence constraints” in accordance with the modern theory of the a priori formation of the system of the measured properties (5, 6) in the FCA methodology. The existence constraints between the IAM dimensions significantly reduce the speed of calculating the concepts and reduce their number in comparison with the control model. In addition, the analysis of the properties of the concepts makes it possible to

identify the boundaries of the size of the extent and intent of the concepts due to the natural limitations of the number of analytical links in real databases.

4 Conclusion

On-line processing of big data requires high-speed performance in the conditions of a large volume and heterogeneity of information. The results of the performed computational experiment show that the representation of the integral analytical model of heterogeneous data as a lattice is suitable for solving modern problems of the real-time analysis of big data. The development of the proposed approach is associated with the systematization of the previously obtained results and a description of the full cycle of creation and use of the integral analytical model. Improving the theoretical basis for the approach consists in intellectualizing the process of forming IAM in terms of building a multidimensional model and taking into account the variability of analytical relationships.

References

1. Lohr, Steve: When There's No Such Thing as Too Much Information. New York Times. Aprill 23 (2011)
2. Abelló, A., Romero, O., Pedersen, T. B., Berlanga, R., Nebot, V., Aramburu, M. J., & Simitsis, A.: Using semantic web technologies for exploratory OLAP: a survey. IEEE transactions on knowledge and data engineering **27(2)**, 571–588 (2014). doi: https://doi.org/10.1109/TKDE.2014.2330822
3. Alpar, P., Schulz, M.: Self-Service Business Intelligence. Bus. Inf. Syst. Eng. **58(2)**, 151–155 (2016). doi: https://doi.org/10.1007/s12599-0160424-6
4. Sivarajah, U., Kamal, M. M., Irani, Z., Weerakkody, V.: Critical analysis of Big Data challenges and analytical methods. Journal of Business Research. **70**, 263–286 (2017). doi: https://doi.org/10.1016/j.jbusres.2016.08.001
5. Korobko, A., Korobko, A.: Multidimensional Design from XML Sources for the Integral Analytical Model. DEStech Trans. Comput. Sci. Eng. AIIE (2017). doi: 10.12783/dtcse/aiie2017/18203
6. Korobko, A., Korobko, A.: Matching disparate dimensions for analytical integration of heterogeneous data sources. In: Proceedings of the 11th International Conference on Management of Digital EcoSystems (MEDES '19). pp. 66–72 (2019). doi: https://doi.org/10.1145/3297662.3365809
7. Pahomova K.I., Korobko A.V.: Application of formal conceptual analysis for intelligent decision support. Artificial Intelligence and Decision Making. **4**, 37–46 (2019) (in Russian)
8. Andrews, S.: In-close, a fast algorithm for computing formal concepts. In: International Conference on Conceptual Structures, Moscow (2009)