UDC 514.18

COMPOSITIONAL GEOMETRIC METHOD OF INFORMATION ANALYSIS AND ITS APPLICATION WHEN WORKING WITH BIG DATA

Vereshchaga V.M., Doctor of Technical Sciences,
vervik49@gmail.com, ORCID: 0000-0003-0038-8300
Adoniev Y.O., Doctor of Technical Sciences,
evgen.adoniev@gmail.com, ORCID: 0000-0003-1279-4138
Melitopol School of Applied Geometry
Zaporizhia National University (Melitopol, Ukraine)

The article proposes a composite geometric method for analysis of information in Big Data sets at the stage of their primary processing and "cleaning". The method is based on the methods of the Baluba-Naydysh point calculus is a preparatory stage when using the structural geometric modelling of Big Data.

For effective analysis of Big Data, it is important to use appropriate sorting algorithms by number in certain clusters (groups). In each cluster, the points of the database have the same (within a certain tolerance for deviation), characteristics-coordinates that define them. Using clusters with a large number of points, you can determine the course of the process, identify trends in its development. Clusters with a relatively small number of points, as a result of the analysis, can be excluded from consideration, as those that do not significantly affect the development of the situation. The representation (of objects) of any database in the form of points that have, in quantity and quality, coordinates, fully correspond to their properties and characteristics, we will call compositional geometrization of Data.

Data properties can be completely different in nature and content. During the geometrization of the database, two coordinate systems are applied simultaneously by the methods of compositional geometric modeling. The first is a three-dimensional coordinate system of object space in which the process flows. In this case, a fourth coordinate is added - this is the change in time. The second is the n-dimensional coordinate system of the parameter space, in which the coordinates of the database elements are determined, the properties and characteristics of each element are parameterized.

The Data geometrization process greatly simplifies the next stage of work - the development of compositional geometric models. In particular, the minimal use of machine resources when working with Big Data significantly reduces the cost of obtaining valuable conclusions and forecasts.

Keywords: Big Data, cleaning, primary processing, point BN-calculus, compositional method of geometric modelling.

Formulation of the problem. Efficient use of big data opens up great opportunities for any business. At the same time, inaccurate or incorrectly processed information can lead to erroneous decisions. In this sense, it is important to "clean up" the data, aimed at identifying errors and inconsistencies.

The solution to the problem of "cleaning" large data sets of different nature and origin from errors, duplication, "noise", as well as their initial analysis will be greatly facilitated by the development of effective methods of working with data. These methods will involve formalizing and comparing information from different large databases.

Existing data cleaning tools are the lack of interactivity when the conversion is performed in a batch process without feedback. Data often have many "nested discrepancies" that are difficult to detect. Both data conversion and mismatch detection require user effort, which complicates the cleaning process and reduces its quality.

With the advent of a new geometric apparatus of point BN-calculus [1] and developed on its basis by the compositional method of geometric modeling [2], it became possible to geometrically analytically formalize large amounts of data of different nature. The advantages of the geometric method of analytical formalization are universality, simplicity of algorithmization and conciseness [3]. Therefore, the development of methods of compositional geometric analysis for the primary processing of large data sets is an urgent problem, the solution of which will allow more efficient, better analysis of large data.

Analysis of recent research. In most cases, the cleaning and initial analysis of large data sets of different nature and origin are based on data profiling and data mining [4]. Methods specially developed for big data are used: Cleanix, SCARE, KATARA, BigDansing [5]. These methods provide a systematic approach to standardization of representation, elimination of duplicates, detection of anomalies and removal in "dirty" databases. They include the following stages: data processing, data processing, validation and verification of data. The methods are quite complex and resource-intensive when working with big data.

The use of a composite geometric method of data analysis, in our opinion, will greatly simplify the work with large data sets, significantly reduce errors and reduce the cost of computer resources in the initial processing of large data.

An approach to the purification of input data through their geometrization is proposed for the first time.

The aim of the study. To increase the efficiency of primary processing of large data sets of different nature, we propose a composite geometric method of big data analysis.

Main part. We use the Balyuba-Naidysh point calculus apparatus as a base for our study. The elements of any database can be represented as points, which are determined by the number of coordinates $k = \overline{1,n}$, either in the object space or in the parameter space.

For effective analysis of big data it is important to use appropriate algorithms for sorting by number in certain clusters (groups). In each cluster, the elements-points of the database have the same (within a certain tolerance for deviation), the characteristics-coordinates that define them. Using clusters with a large number of points, you can determine the course of the process, identify trends in its development. Clusters with a relatively small number of points, as a result of the analysis, can be excluded from consideration as those that do not significantly affect the development of the situation.

Representation of elements (objects) of any database in the form of points that have, in quantity and quality, coordinates that fully correspond to their properties and characteristics, will be called composite geometrization of data. Data properties can be completely different in nature and content.

Note that each element of any database has properties and characteristics that may differ in their parameters from other related elements of the same database. Accordingly, the points corresponding to these elements (which are obtained as a result of geometrization of these elements) will have different numbers of coordinates in the n-dimensional space of parameters, where these characteristics are geometrized.

Composite geometric modeling (CGM) involves, when creating a composite model, the simultaneous use of points with different numbers of coordinates.

Therefore, the different number of coordinates at the points that geometrize the elements of the database is not an obstacle to the creation of models by the methods of composite geometric modeling. Unlike CGM, existing methods of geometric modeling require the same number of coordinates at the basis points. The base points discretely, in the form of the original geometric figure, represent the course of the simulated process.

During the geometrization of the database by the methods of composite geometric modeling, two coordinate systems are used simultaneously. The first is the three-dimensional coordinate system of the object space in which the process takes place. In this case, the fourth coordinate is added - it is a change of time. The second is an n-dimensional coordinate system of the parameter space, which determines the coordinates of the database elements, which parameterize the properties and characteristics of each element.

These two coordinate systems of the object three-space and the n-space of parameters are interconnected through the application of a simple relationship of three points. This relation is the basis for both the point calculus of Balyuba-Naidysh and for compositional geometric modeling. In this case, changes or even replacements of points in one coordinate system do not entail any changes in another coordinate system. If it is necessary to replace the individual points of the input geometric figure, which entail changes in both coordinate systems, the created, composite geometric model remains unchanged. This is possible due to the fact that in the formula records of composite geometric models, the basis points are always recorded separately from the parameter records, which ensure the continuity of the composite geometric model.

The process of data geometrization greatly simplifies the next stage of work - the development of composite geometric models.

Here is an example of one of the possible options for geometric analysis of the input database. Let some database be geometrized, as a result of which we have a finite set of points containing millions of points - N. Depending on the course of the studied process continuously, or on each of its segments, the maximum N_{max} and minimum N_{min} points are determined (Fig. 1).

There is a point O - the center of concentric circles:

$$O = \frac{N_{max} + N_{min}}{2}.$$

Figure 1 shows three circuits. However, depending on the requirements for database analysis, such circles may be more or less. As you can see (Fig. 1), the center O is on the line segment (N_{max} N_{min}). The points N_{max} and N_{min} can be located anywhere on the outer circle.

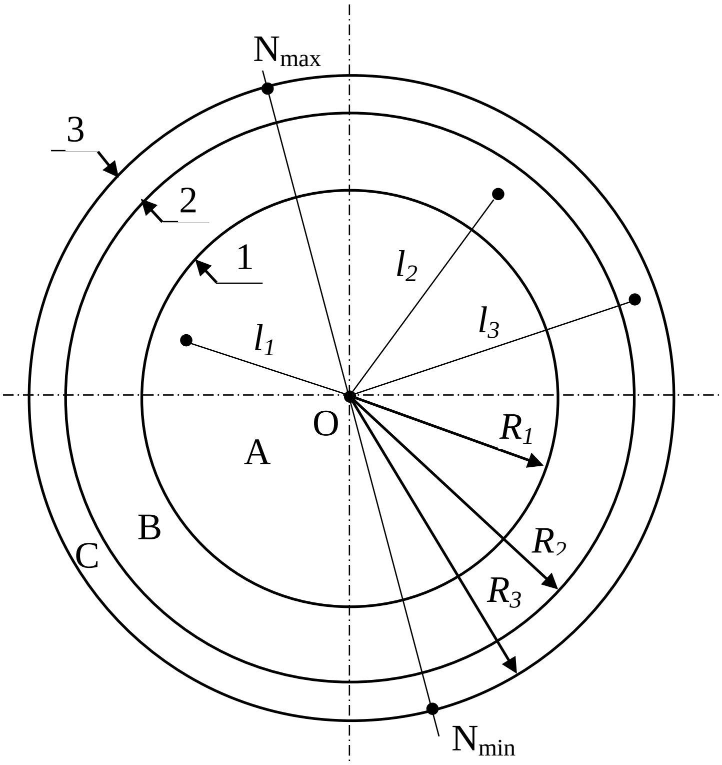

Fig.1. Scheme for a geometric way to remove erroneous data and noise

The three circles -1, 2, 3 are divided into all database points into three classes, which are detected by comparing the distance l from the center to the current point with the length of the radius R of the corresponding circle.

For all points of class A there should be $l_1 \leq R_1$, for all points of class B there should be $R_1 < l_2 \leq R_2$ and for all points of class C there should be $R_2 < l_3 \leq R_3$.

The outer circle has R_3 = const. The radii R_1 and R_2 can be changed, respectively, in the range: $0 \leq R_1 < R_2$; $R_1 < R_2 < R_3$. By varying the radii R_1 and R_2, you can change the number of points in classes A and B. Such changes are made based on the requirement to analyze the input data. Reducing the difference R_3-R_2, we can determine the class C, in which only erroneous data and noise components will remain. Noise includes random objects that have been captured and entered into a database.

Note that for each date set you need to develop a separate method of geometrization of its elements, which is a difficult task. In some cases, it is most effective to apply the method of geometrization of input data immediately at the stage of formation of the primary database. As an example, the signal from the sensor can enter the data warehouse in a geometrized form.

Therefore, data processing using a composite geometric method of information analysis, when database elements are presented in the form of multidimensional points (in terms of composite geometric modeling), is quite efficient because it is carried out by simply comparing modules of distances from center to current points and radius , which defines the boundaries of the class. The geometric method of compositional analysis eliminates the use of complex algebraic operations, which are part of the algebraic methods of input data analysis, and which are always more resource-intensive than geometric methods. This is especially important when working with large data sets, where millions of iterations are performed at each stage of the analysis. Therefore, the effectiveness of the calculation methods used is a key success factor in big data.

Another feature of working with big data is their heterogeneity. Often the data comes from different sources, have different physical nature and meaning. The data format is not only numerical but also, in particular, text, graphic, video, audio, etc.

The real value of big data results directly depends on the integration of different types of data sources and their large-scale analysis. Combining different data streams has a synergistic effect, already significantly increases their value, even before their analysis.

Methods of compositional geometric analysis make it possible to geometrize heterogeneous database elements. In addition, composite geometric models can simultaneously contain the basis points of the input geometric figure with a different number of coordinates of the n-dimensional space of parameters obtained by geometry of heterogeneous database elements.

Under these conditions, the geometrization of input data of any nature and from any source, leads the data to a single format in terms of the proposed compositional analysis. Such geometrized data is convenient to analyze, build models using the compositional method of geometric modeling. For a large enterprise or organization, this makes it possible to most effectively find non-obvious dependencies in the arrays of big data, make predictions and provide information-based recommendations for appropriate management decisions.

Thus, the application of composite geometric analysis of information will have the maximum economic effect when applied in the field of data science in general, and in the field of big data in particular.

Conclusions. A new approach to the primary processing of large data sets of different nature is proposed, namely: composite geometric method of big data analysis. It involves the representation of the elements of the input database in the form of multidimensional points of the n-space of parameters. An example of an algorithm for applying a geometric method to remove erroneous data and noise is given.

The application of the developed geometric method at the preparatory stage of development of composite geometric models is perspective.

The economic effect of the introduction of the proposed method of preparation and "purification" of data is based on the advantages of the compositional method of geometric modeling. In particular, the minimal use of machine resources when working with big data significantly reduces the cost of obtaining valuable conclusions and forecasts. This makes them available to a wide range of relatively small businesses.

Literature

1. Balyuba I.G., Naidysh V.M. Point calculus [textbook]. Melitopol: MDPU Publishing House. Bogdan Khmelnytsky, 2015. 234 p.
2. Adoniev Y.O. Compositional method of geometric modeling of multifactor systems: dis. ... Dr. Tech. Science. К .: KNUCA, 2018. 512 p.
3. Vereshchaga V.M. Composite geometric modeling: Monograph. Melitopol, 2017. 108 p.
4. Poltavtseva M.A., Zegzhda D.P., Kalinin M.O. Big data management system security threat model. *Automatic control and computer sciences.* 2019. No 8. P. 903–913.
5. Zageeva L.A., Nizamov T. The main directions of big data in the banking sector. *Innovative economy and law*. 2019. No. 2 (14). P. 40- 44.

КОМПОЗИЦІЙНИЙ ГЕОМЕТРИЧНИЙ СПОСІБ АНАЛІЗУ ІНФОРМАЦІЇ ТА ЙОГО ЗАСТОСУВАННЯ ПРИ РОБОТІ З ВЕЛИКИМИ ДАНИМИ

Верещага В.М., Адоньєв Є.О.

У статті запропоновано композиційний геометричний спосіб аналізу інформації у великих дата сетах на етапі їх первинної обробки та «очищення». Спосіб базується на методах точкового числення Балюби-Найдиша та є підготовчим етапом при застосуванні композиційного методу геометричного моделювання великих даних. Для ефективного аналізу великих даних важливе застосування відповідних алгоритмів сортування за кількістю у певних кластерах (групах). У кожному кластері *елементи-точки бази даних мають однакові (у межах визначеного допуску на відхилення), характеристики-координати, що їх визначають. Використовуючи кластери з великою кількістю точок, можна визначити перебіг процесу, виявити тренди його розвитку. Кластери з відносно невеликою кількістю точок, в результаті аналізу, можуть бути виключені з розгляду, як такі, що не впливають суттєво на розвиток ситуації. Подання елементів (об'єктів) будь-якої бази даних у вигляді точок, які мають, у кількості та якості, координати, що у повній мірі відповідають їх властивостям та характеристикам, будемо називати композиційною геометризацією даних. Властивості даних можуть бути геть різними за суттю та змістом. Під час геометризації бази даних методами композиційного геометричного моделювання застосовуються одночасно дві системи координат. Перша – це тривимірна система координат об'єктного простору, у якому відбувається перебіг процесу. При цьому, додається четверта координата – це зміна часу. Друга – це n-мірна система координат простору параметрів, у якій визначаються координати елементів бази даних, які параметризують властивості та характеристики кожного елементу. Процес геометризації даних набагато спрощує наступний етап роботи – розробку композиційних геометричних моделей. Зокрема, мінімальне використання машинного ресурсу при роботі з великими даними значно здешевлює отримання цінних висновків та прогнозів.*

Ключові слова: великі дані, очищення, первинна обробка, точкове БН-числення, композиційний метод геометричного моделювання.

КОМПОЗИЦИОННЫЙ ГЕОМЕТРИЧЕСКИЙ СПОСОБ АНАЛИЗА ИНФОРМАЦИИ И ЕГО ПРИМЕНЕНИЕ ПРИ РАБОТЕ С БОЛЬШИМИ ДАННЫМИ

Верещага В.М., Адоньев Е.А.

В статье предложен композиционный геометрический способ анализа информации в больших дата сетах на этапе их первичной обработки и «очистки». Способ базируется на методах точечного исчисления Балюбы-Найдыша и является подготовительным этапом при использовании композиционного метода геометрического моделирования больших данных. Для эффективного анализа больших данных важно применение соответствующих алгоритмов сортировки по количеству в определенных кластерах (группах). В каждом кластере элементы-точки базы данных имеют одинаковые (в пределах определенного допуска на отклонение), характеристики-координаты, которые их определяют. Используя кластеры с большим количеством точек, можно определить ход процесса, выявить тренды его развития. Кластеры с относительно небольшим количеством точек, в результате анализа, могут быть *исключены из рассмотрения, как такие, которые не влияют существенно на развитие ситуации. Представление (объектов) любой базы данных в виде точек, которые имеют, в количестве и качестве, координаты, в полной мере соответствуют их свойствам и характеристикам, будем называть композиционной геометризацией данных. Свойства данных могут быть совершенно разными по сути и содержанию. Во время геометризации базы данных методами композиционного геометрического моделирования применяются одновременно две системы координат. Первая - это трехмерная система координат объектного пространства, в котором происходит течение процесса. При этом, добавляется четвертая координата - это изменение времени. Вторая - это n-мерная система координат пространства параметров, в которой определяются координаты элементов базы данных, параметризируют свойства и характеристики каждого элемента. Процесс геометризации данных намного упрощает следующий этап работы - разработку композиционных геометрических моделей. В частности, минимальное использование машинного ресурса при работе с большими данными значительно удешевляет получение ценных выводов и прогнозов.*

Ключевые слова: большие данные, очистка, первичная обработка, точечное БН-исчисление, композиционный метод геометрического моделирования.

Referenses

1. Balyuba, I.G., Naidysh, V.M. (2015) *Point calculus*/ In V.M. Vereshchaga. Melitopol: MDPU name Bogdan Khmelnytsky [in Ukrainian].
2. Adoniev, Y.O. (2018) *Compositional method of geometric modeling of multifactor systems*. Doktor's thesis. Kiyev: KNUCA [in Ukrainian].
3. Vereshchaga, V.M. (2017). *Composite geometric modeling*: Monograph Melitopol: FOP Odnorog T.V. [in Ukrainian].
4. Poltavtseva M.A., Zegzhda D.P., Kalinin M.O. (2019) Big data management system security threat model. *Automatic control and computer sciences*. 8. 903–913. [in Netherlands].
5. Zageeva L.A., Nizamov T. (2019). The main directions of big data in the banking sector. *Innovative economy and law*. 2019. 2 (14). 40- 44. [in Singapore].