Models of storage and preliminary processing of retrospective data for the energy company

Victor, Potapov

Omsk State Technical University
Omsk, Russia
e-mail: ivt@omgtu.ru

Igor, Chervenchuk

Omsk State Technical University
Omsk, Russia
e-mail: cherven-igor@yandex.ru

Vladislava, Samus

Omsk State Technical University
Omsk, Russia
e-mail: sva.1994@yandex.ru

Aleksandr, Gritsay

Omsk State Technical University
Omsk, Russia
e-mail: aleksandr.gritsay@gmail.com

ABSTRACT

At present, for the entity of the electricity market - an energy sales company - one of the main business processes is the process of solving the problems of short-term forecasting of power consumption. In this case, the expert, as a rule, has to work with a large array of retrospective data on power consumption and their samples. Such array of data has a simple homogeneous structure. This raises the problem of combining data from various sources and it takes a long time to obtain various data samples, and therefore automation of the preliminary processing of data for subsequent statistical analysis in conditions of a large amount of initial information on power consumption is relevant.

Keywords

Computer science, informatics, information technology.

1. INTRODUCTION

When constructing prognostic models of electricity load [1, 2], the subject of the Electricity Market (EM), as a rule, operates with the data of automated commercial accounting systems.

For the primary storage of retrospective data on electricity load, the data model described in terms of objects and indicators can be used successfully [3]. The proposed model defines the objects-carriers of properties, described by means of a system of indicators. In general, it is assumed that the description object is structured, consisting of conceptually homogeneous elements that are hierarchically linked. The system of indicators is also partially or completely structured, hierarchical. Thus, two hierarchies are distinguished: a hierarchy of objects and a hierarchy of indicators (properties) used to describe retrospective data on electricity load. Further, the model in question will be called the indicator-object model, the main purpose of which is to provide convenient storage of structured data on the electricity load of consumer objects for subsequent statistical analysis. Application of such an approach in the description of data makes it possible to present in a natural and compact form data for storage and use.

2 RESEARCH

Fig. 1 in the UML notation shows the

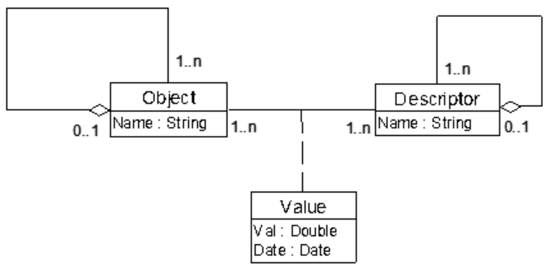

Fig. 1. Structure of the model Indicator-object

data structure of the system, built on the basis of two hierarchies: the hierarchy of objects, and a hierarchy of indicators, oriented to conceptually homogeneous data, described by numerical values. The first hierarchy - the hierarchy of objects is defined by the reflexive aggregation of the Object class, which assumes a "part-whole" type relationship between the elements of the structured object, for example, the chain "Consumer, consumption object, accounting unit, meter". The second hierarchy, the indicator system, is defined by the reflexive aggregation of the Descriptor class, which assumes a structured description of the indicator, for example, the electricity load of the object. There is a fundamental difference between the hierarchy of objects and indicators that any object, regardless of its place in the hierarchy of objects, can have a numerical value, whereas only a fully defined indicator - an element of the end of the hierarchy, has specific values.

The actual value is not mapped separately with any indicator and not with any specific object, but with a pair of indicator object, so the Value class is an association class between objects and indicators. The presence of the Date attribute, which contains the date and time of the measurement of power consumption at the given object, allows later to sort data by dates, perform data analysis, and build a language for manipulating data with wide possibilities [5].

The hierarchy is a natural and convenient means of searching for ordered data, the basis for constructing a searchable hierarchical menu that allows the expert to easily navigate to the final data of the dynamically formed pricing model.

The considered data storage model is easily displayed on the relational data model.

Hierarchical structures of indicators (x) and objects (y) with the corresponding values (z) and dates of their measurements (d) specify a four-dimensional information retrieval space (IR space) with a number of special properties. The concept of IR space allows us to formalize the procedures for preliminary preparation of data for pricing analysis.

The use of the proposed apparatus, which we call the apparatus for calculating samples, allows us to formalize the problem of preliminary preparation of the price data for constructing the rank distribution. For this purpose, the time reference of the initial data is used, allowing to accumulate data on the pricing model with the possibility of taking into account its dynamics and changes. In our case, a sample will be called a time-ordered set of values for all time moments of a certain indicator at some object. The sample contains data. We will call the pair <date, value>. Thus, the sample is an ordered set (sorted by increasing the date / time attribute sequence) of the data. In this context, the sample can be represented by a table of two columns (Table 1).

Table 1

example of sample

Date	Value
1.01.2015	12,1
1.02.2015	13,2
1.03.2015	11,7
1.04.2015	11,2

Introduce the sample addition operation (+). Let us have two objects of the same rank, the results of observing the object from each source of observations form samples v1 and v2, respectively. The selection $v3$ – Amount of samples $v1 \overset{\cup}{+} v2$ can be found by adding the sample v1 to the data from v2, having dates not in v1, if the data v1 and v2 have equivalent dates, then the average value for v1 and v2 is entered in v3.

As the equivalence of dates, you can use equality or belonging to a certain period (week, decade, month, etc.)

Obviously, the addition of samples is commutative, idempotent, but associativity is performed only when data with equivalent data occurs no more than once. This operation on the samples is determined by the corresponding operation on the data.

There are 0 operations - empty sample - v0, which does not contain any given data (or you can consider that all data contains NULL-values). For any sample v1

$$v0 \overset{\cup}{+} v1 = v0 \overset{\cup}{+} v1 = v1 \qquad (1)$$

The operation of addition of samples can also be applied to different objects and indicators, then this operation will make sense of generalization.

However, it is not possible to enter an operation such as a sample difference.

Thus, on the set of admissible samples V we have the algebra A + with the two-place operation of addition of samples and the null operation-an empty sample. This algebra A + is an idempotent commutative group id with unit and in this form, is not very convenient for use. one. Thus, the extended sample is an ordered set (sorted by increasing date / time sequence) of the data. In this context, the sample can be represented by a table of three columns, as shown in Table. 2

Initially, for all sample data, the degree of averaging is 1.

Table 2

Example of an extended sample

Date	Value	G
1.01.2015	12,1	1
1.02.2015	13,2	2
1.03.2015	11,7	1
1.04.2015	11,2	1

On the set of extended samples V_, we introduce the addition operation ($\overset{\cup\bullet}{+}$) – more precisely, the operation of adding to the union of extended samples. Supposably $v3=v1 \overset{\cup\bullet}{+} v2$. If for some data from v1 and v2 the dates are equivalent,

then $\begin{cases} v3.g = v1.g + v1.g; \\ v3.val = \dfrac{v1.val \cdot v1.g + v2.val \cdot v2.g}{v1.g + v2.g}. \end{cases}$ (2)

If the dates do not match, i.e. Some given data with the date / time d of one operand, there is no given equivalent date in the other operand (or there is a NULL-data), then the values are transferred to the result. Here the averaging number of the given shows how many values were involved in averaging. Table 5 shows the sum of the extended samples from 3 and 4.

Table 3

Sample of v1

Date	Value	G
1.01.2015	10,2	1
1.03.2015	12,1	1
1.05.2015	13,4	2
1.06.2015	11,7	1
1.07.2015	11,2	1

Table 4

sample of v2

Date	Value	G
1.01.2015	10,8	1
1.02.2015	4,0	2
1.04.2015	7,0	1
1.05.2015	13,2	2
1.06.2015	11,5	1

Table 5

Sample of $v3 = v1 \overset{\cup\bullet}{+} v2$

Date	Value	G
1.01.2015	10,5	2
1.02.2015	4,0	2
1.03.2015	12,1	1
1.04.2015	7,0	1
1.05.2015	13,3	4
1.06.2015	11,6	2
1.07.2015	11,2	1

The sum of a large number of samples will tend to the vector of mathematical expectations of a certain indicator for different dates. Obviously, this operation is associative, commutative, but it is not strictly idempotent. The algebra formed by the operation $\overset{\cup\bullet}{+}$ on V_ is a commutative semigroup (with respect to the zero of the operation-an empty sample-will be discussed later).

On extended samples, it is possible to conditionally set the operation opposite to the summation - the difference $\overset{\cup\bullet}{-}$ (More precisely, the difference between the extended samples for the union).

Let $v3=v1 \overset{\cup\bullet}{-} v2$. If for some data from v1 and v2 the dates are equivalent,

then $\begin{cases} v3.g = v1.g - v1.g; \\ v3.val = \dfrac{v1.val \cdot v1.g - v2.val \cdot v2.g}{v1.g - v2.g} \end{cases}$ (3)

If for some date / time d1 the given type <d1, z1, g1> is only in one sample, then we assume that in the other sample there is a NULL-given of the type <d1, $\frac{0}{0}$, 0> and apply (3). The use of uncertainty of such a $\frac{0}{0}$, in this case it is permissible, since in the subsequent calculations this uncertainty is removed, since in both (2) and (3) the value of the given (v.val) is always multiplied by the degree of averaging (v.g), i.e.

$$v.val \cdot v1.g = \frac{0}{0} \cdot 0 = 0.$$

Thus, for extended samples, NULL data has a clear interpretation as data of the form <$d1$, $\frac{0}{0}$, 0> (where $\frac{0}{0}$ - NULL value), the result of the sum and difference of some given with NULL-data equals this given.

If we add the sample v1, presented in Table. 3, NULL-data and subtract it from the sample v3 is presented in Table 2, in a result get a sample of v2, which was previously presented in 4, supplemented with NULL-data.

Operations on extended samples are defined by the corresponding operations on the extended data, the zeros of these operations are NULL data. It is assumed that each sample contains the required amount of NULL data (this number depends on the data in other samples). NULL-sampling (v0) is represented by the necessary number of NULL-data, all NULL-samples are equivalent and meaningful are empty. This approach allows you to perform addition and difference operations, directly using only formulas (2) and (3) for all relevant data.

When using operations $\overset{\cup\bullet}{+}$ and $\overset{\cup\bullet}{-}$ in an arbitrary order, fictitious data may appear with averaging degree g≤0, which in subsequent operations will be eliminated. Fictitious data is not involved in external operations, but is used only for sample manipulation. NULL-given is fictitious.

Thus, on the set of extended samples V_ we have the algebra A +/- with the two-place operation of addition of samples, the single-step operation of taking the opposite element and the null operation-an empty sample. This algebra A +/- is an abelian group.

For source samples, there is a single-valued mapping $h^{-1}:V\rightarrow V\_$ in which for all data (except implicitly present fictitious ones), 1 (unit) is entered in the averaging attribute.

The calculations on the samples can be performed as follows: first, we move from the initial samples to the extended samples using the operations on the extended samples, perform the calculations, and convert the final result by h to the usual sample.

When performing complex long-time calculations, intermediate results should be stored in an expanded format, the original data and results in the usual.

Over samples – operations of comparison and co-location, allowing to represent the entire spectrum of statistical calculations, which can serve as the basis for creating a kind of language for data manipulation.

3. CONCLUSION

The attractiveness of the offered approach consists, first of all, in that it allows to formalize and automate the processes of preparation and preliminary processing of data of billing systems of energy sales companies, to compile the required data model, for the purpose of subsequent statistical analysis. The offered model was successfully used to automate the processes of preliminary processing and data storage when solving short-term forecasting of electricity load problems and formed the basis for the developed program for short-term forecasting of electricity load [6].

The offered model can also be used for preliminary processing, storage and analysis of primary information on

This supplement (i.e., a pair of <date-set values>), additional attribute (G), defined on the set of integers, and comprising averaging this number, this will be called the extended data. The expanded data will be called the "date", "value", "number-average". The number of averaging shows how much data was involved in obtaining the value as an average for the date equivalence, initially for all data of the sample the averaging degree is equal to external factors, for example, meteorological such as temperature, illumination, which provides a methodological basis for constructing more accurate electricity load forecasts that take into account the influence of external factors.

REFERENCES

[1] A. S. Gritsay and D. A. Tyunkov, “Classification methods for predicting the short-term electricity OREM subjects,” (in Russian), in *Proc. All-Russia Workshop on Actual energy issues*, Omsk, Russia, 2016, pp. 41-45.

[2] V.I. Potapov, *et al*., “Using neural network to construct a short-term forecast of electricityconsumption LLC “Omsk Energy Retail Company”,” (in Russia), in*Izvestiyatomskogopolitekhnicheskogouniversiteta. Inzhiniringgeoresursov*, vol. 327, no. 8, pp. 44-51, 2016.

[3] I. V. Chervencuk, “Model presentation of statistical data,” (in Russia), in *Proc. 19th Int. Workshop on Natural and intellectual resources of Siberia (SIBRESURS-19-2013)*, Tomst, Russia, 2013, pp. 187-191.

[4] I. V. Chervencuk, “Characteristics efficiency indicators of structure,” (in Russia), in *Omskiynauchnyyvestnik*, no. 20, pp. 149-151, 2002.

[5] I. V. Chervencuk, “Language manipulate data processing system of statistical information on the basis of sample calculation,” (in Russia), in *Proc. 4th Int. Workshop on Dynamics of Systems, Mechanisms and Machines*, Omsk, Russia, 2002, vol. 1, pp. 245-247.

[6] R. N. Khamitov, A. S. Gritsay and D. A. Tyunkov, “Program of short-term forecasting of electricity consumption using the artificial neural network for wholesale electricity and capacity market energy retail companies,” (in Russia). Moscow: OFERNiO, 2016.The certificate of registration of an electronic resource N 22122 on September 1, 2016.