ALGORITHM BIDIMS FOR AUTOMATED SYSTEMATIZATION OF DATA ARRAY. CASE STUDY: REDISCOVERING MENDELEEV'S PERIODIC TABLE OF CHEMICAL ELEMENTS.

Andrey Zagoruiko and Nikolay Zagoruiko

Abstract: The method (algorithm BIDIMS) of multivariate objects display to bidimensional structure in which the sum of differences of objects properties and their nearest neighbors is minimal is being described. The basic regularities on the set of objects at this ordering become evident. Besides, such structures (tables) have high inductive opportunities: many latent properties of objects may be predicted on their coordinates in this table. Opportunities of a method are illustrated on an example of bidimentional ordering of chemical elements. The table received in result practically coincides with the periodic Mendeleev table.

Keywords: bidimensional structure, data mining, ordering, prediction, approximation.

Introduction

One of Data Mining purposes consists in reduction of the available data to such kind at which the person easily perceives the basic contents of the analyzed information. The information advanced in such a way becomes a little more substantial for the person. The specified purpose is achieved by different means. The important role plays machine graphics, allowing the person to perceive the information through the powerful visual analyzer. Various ways of ordering at which one-dimensional or bi-directional data file will be coordinated well with the simple concepts (models) have a wide circulation. One-dimensional line of smoothly growing or decreasing numerical values is evident, for example. Even more information is contained in bi-directional tables with monotonic change of data on the first and second coordinate. It is not surprising that many fundamental laws of a nature - the law of the Ohm, Newton, Mendel etc. are well illustrated by bi-directional data tables.

D.I. Mendeleev, studying dependencies between various properties of chemical elements, relied on many results of the predecessors. In particular, the grouping of some elements on a generality of their chemical properties was known, on similarity of nuclear weights etc. Mendeleev put the task before itself to make such bi-dimensional ordering of all 63 elements known at that time at which the neighbouring elements of the table would be similar to each other on maximum big set chemical and physical properties. Such their arrangement will be coordinated with the concept of local smoothness, which provides easy perception of the general laws for the table.

The use of concepts beforehand prepared or the models is useful not only for an explanation of the data analysis result, but also for the process of this result receiving. The rich history of scientific discoveries speaks about it, in many of which traces of attempts of such type are obviously visible: " - It seems these objects have a S-structure. And what if to try use such a model? And what if to order the objects by such rule?"

Let's illustrate the utility of the initial empirical data association with the simple concepts during revealing the laws latent in data. We'll do it on an example of automatic rediscovering of the periodic law of chemical elements. Such attempt was already done by us [1], but in a little bit idealized conditions. In particular, true nuclear weights and valences on hydrogen, instead of those known to Mendeleev were used; the program was adjusted to the fixed number of properties etc. In the given work those data and knowledge, which D.I. Mendeleev had at his disposal during the creation of the periodic law of chemical elements, are used.

Description of Systematization Algorithm

Let us imagine that we have some data array, consisting of *n* elements *A*, each of them characterized by a set of *k* properties. In the current formulation the task is to systematize this array in a form of a two-dimensional table with internal uniformity in changing of element properties in both dimensions.

Let us also propose that we have some «embryo» of the table, i.e. credible group of small amount of elements, constructed on the base of researcher intuition. Having such a group it is possible to try to predict the properties of neighbouring elements. Assuming the uniformity of elements properties changing inside the «embryo» in both

dimensions, this procedure may be performed using linear approximations. Possible types of such approximations are given below:

Internal approximation may be applied if it is necessary to predict the properties of the element, situated between two elements provided that their positions in the table are already known. In this case if known elements are situated in the table at coordinates, for example, *(i+1,j)* and *(i-1,j)*, then the value of *m-th* property in position (i,j) P_{ij}^{m} may be predicted using the equation

$$P_{ij}^{m} = \frac{P_{i-1\,j}^{m} + P_{i+1\,j}^{m}}{2} \tag{1}$$

Similar estimations may be obtained as well for combinations *(i+1,j+1)* и *(i-1,j-1), (i+1,j-1)* и *(i-1,j+1), (i,j+1)* и *(i,j-1)*, i.e. via the horizontal, vertical and two diagonals (total – 4 variants).

External approximation is used when it is necessary to construct the forecast for the table cell adjacent to the pair of situated elements with known positions. For example, if known elements are placed in the cells with coordinates *(i-2,j)* and *(i-1,j)*, then value of P_{ij}^{m} may be predicted as follows:

$$P_{ij}^{m} = 2P_{i-1\,j}^{m} - P_{i-2\,j}^{m} \tag{2}$$

Here the predictions also can be made via horizontals, verticals and diagonals (total – 8 variants).

Corner approximation is applied when known elements are placed in the table in form of "corner", for example, in the cells with coordinates *(i+1,j), (i,j+1)* and *(i+1,j+1)*. In this case it is necessary to use the equation

$$P_{ij}^{m} = P_{i+1\,j}^{m} + P_{ij+1}^{m} - P_{i+1\,j+1}^{m} \tag{3}$$

Four variants of "corner" positions are possible. The final prediction of the property P_{ij}^{m} is defined as averaged value of all forecasts according to equations, which total number may reach 16.

The next step of the procedure is selection of optimal "pretending" element from the set of remaining elements, which are not positioned in the table. The running through of all remaining elements in relation to every empty cell of the table is made with definition of the positioning quality of every element/cell combination. The modulus of deviation between predicted and real element property values may be used as quality criterion:

$$X_{ij}^{h} = \sum_{m=1}^{k} abs(P_{ij}^{m} - R_{h}^{m}) \tag{4}$$

where X_{ij}^{h} - quality criterion of *h-th* element in the cell with coordinates *(i,j),* R_{h}^{m} - real value of *m-th* property for this element. After running of every elements versus every table cell, the table is filled with only one element, which at all set of *h, i* and *j* is characterized with minimum value of X_{ij}^{h}. This procedure is repeated until completion of positioning of all initial elements.

The problem may be complicated by two factors. Firstly, range of parameter values may be significantly different for different properties, resulting in different contribution of each property to the value of X_{ij}^{h} criterion and, therefore, leading to their «inequality of rights». Secondly, it is not evident that every element is described by a full set of properties, so property array actually may include missing values. Correspondingly, the reversed situation is possible as well, when the definite element property is really present, but cannot be predicted, because it is not present in the property sets of elements used for prediction.

First complication is easily solved by normalization of data for each of properties. In the second case it becomes necessary to define for each cell and each element how many properties area really predicted and then normalize criterion X_{ij}^{h} value as follows:

$$X_{ij}^{h} = \frac{\sum_{m=1}^{k} abs(P_{ij}^{m} - R_{h}^{m})}{Z_{ij}^{h}} \tag{5}$$

where Z_{ij}^{h} - number of coincidence «successfully predicted property» / «presence of that property in the element property set» under attempt to place the *h-th* element in the cell with coordinates *(i,j)*. Moreover, as it was demonstrated by test calculations, to improve systematization quality it is better to give preference to elements with higher value of Z_{ij}^{h}, as more reliably determined. Such preference may be realized by different ways, but in this work we used empirical method, based on application of Z_{ij}^{h}, raised to a power higher than one. Particularly, the optimal order value was found to be 3, i.e equation (5) was transformed into:

$$X_{ij}^{h} = \frac{\sum_{m=1}^{k} abs(P_{ij}^{m} - R_{h}^{m})}{(Z_{ij}^{h})^{3}} \tag{6}$$

Systematization of a Full Set of Chemical Elements

As a case study we used a complete set of chemical elements (see Fig.1), e.i. the essential aim of the work was reopening of periodic law, discovered by Dmitry Mendeleev in 1869.

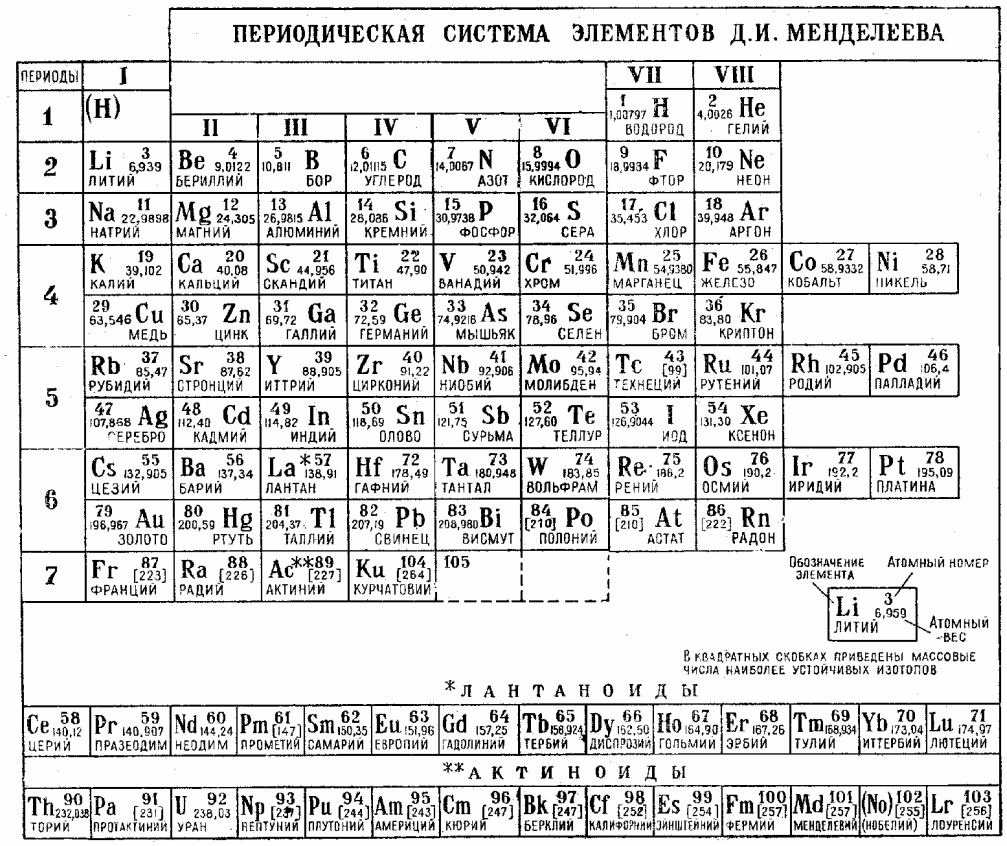

Fig.1. Short-period table of elements

At the first stage we performed test calculations with application of full set of chemical elements, known at the current moment. In case when the set of three basic properties (atomic mass, group number and period number) the described algorithm provided fast and correct solution. Of course, application of group and period numbers

was equivalent to inclusion of already known correct solution into initial data. Therefore, this variant was for software testing purposes only.

On the second stage we used a set of three basic properties, that were known to Mendeleev and which were used by him in his work on periodic law construction: atomic mass, oxygen and hydrogen valence. As "embryos" we used intuitive combinations which look, nevertheless, quite obvious from chemical point of view, such as:

Na			O	F
K	Ca	or	S	Cl

In this case it was possible to successfully construct the "framework" of the table, where periodicity and uniformity of properties changing are evident, namely – 2nd and 3rd periods. The following table filling met significant complications, mainly for transitional element and, especially, for triads Fe-Co-Ni, Ru-Rh-Pd, Os-Ir-Pt and inert gases. Correct positioning of lanthanides and actinides was found to be completely impossible.

At the same time some interesting regularities were discovered. It was found that even one erroneous positioning of an element leads to the chain of further errors, the sooner the error is made the more significant distortions are contributed to the final result. It was also detected that table construction quality (quite logically) depends very much upon the choice of initial "embryo".

Explanations of all these problems are rather simple. Periodicity and uniformity of changing of atomic mass looks evident (Fig.2), at least if will not consider natural mass gap in the area of lanthanides placement (this gap is absent in a long-period table). At the same time the picture for oxygen and hydrogen valences is much more complicated (Figs. 3 and 4).

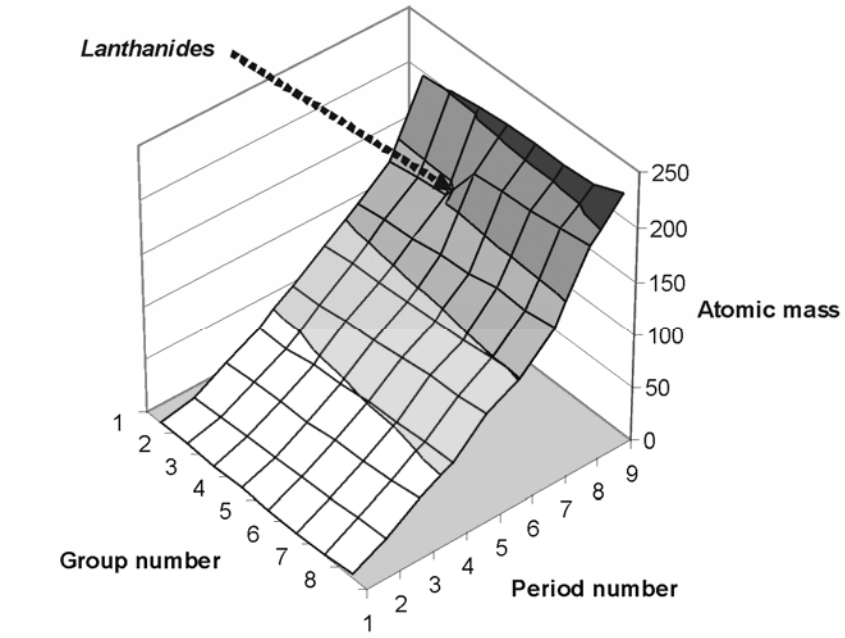

Fig.2. Changing of atomic mass of elements in groups and periods of "short" periodic table.

Good periodicity is observed for hydrogen valence (Fig.4), but this data is present for less than a half of elements. Furthermore, continuity of data is present in 2nd and 3rd periods only, and in higher periods the periodicity is broken by transitional elements (i.e. is repeated "a string after").

Majority of oxygen valence data (Fig.3) is fit into irreproachable flat plane, but with significant anomalies at the table periphery, notably:

- decrease of observed valence in triads Fe (6+) - Co(3+) - Ni(2+), Ru(8+) – Rh (4+) – Pd(2+), Os (8+) – Ir(4+) – Pt(4+);
- zero valence for inert gases (except Xe(8+) and Kr(2+));
- high valences of copper (2+) and gold (3+) instead of expected (1+).

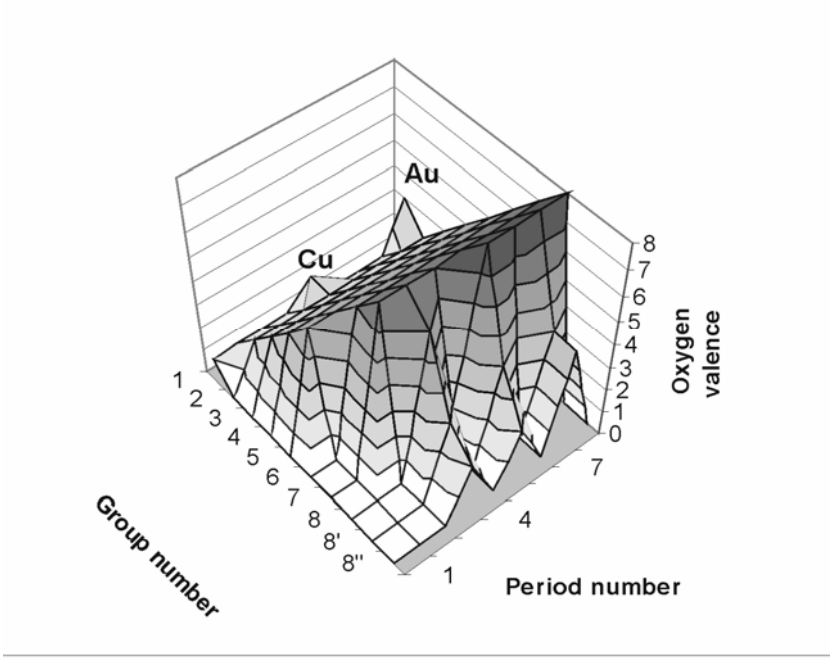

Fig.3. Changing of oxygen valence.

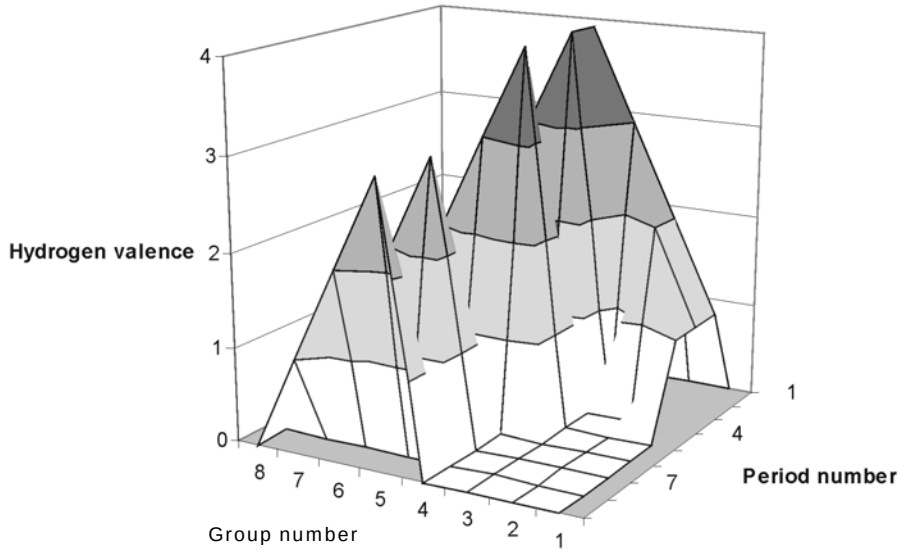

Fig..4. Changing of hydrogen valence.

In general it may be stated that in relation to mentioned properties the data array in Mendeleev's table is not uniform, as it was proposed at the stage of problem formulation. Nevertheless, the attempts were made to "smooth" these nonuniformities by application of greater number of properties in element descriptions. The properties that were definitely known to Mendeleev during his work on Periodic Law (densities, melting and boiling temperatures for elements and their oxides and chlorides; acid/base properties of oxides etc) were chosen to expand data array.

Surprisingly, this attempt was even less successful. The reason for this fault may be demonstrated on the base of element density changing (Fig.5).

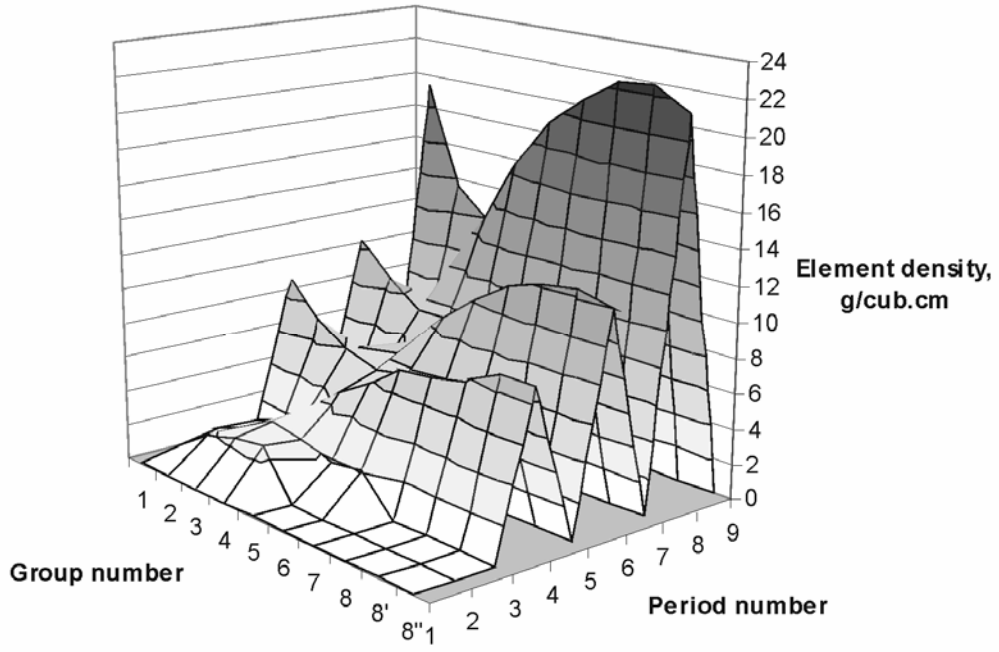

Fig.5. Changing of density of elements in groups and periods of "short" periodic table.

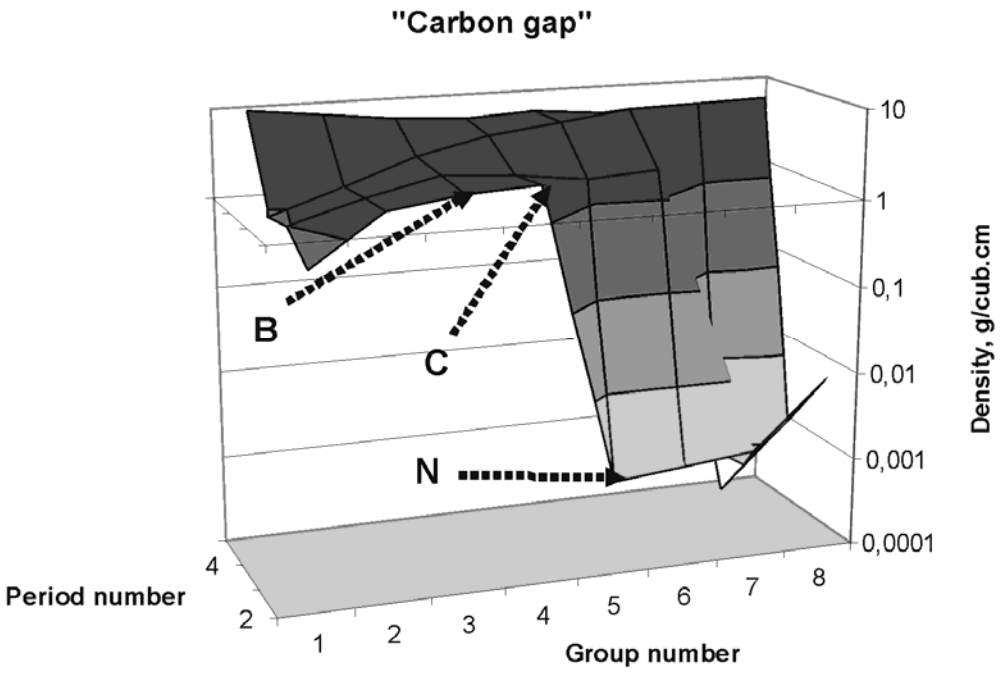

Fig.6. Elements density gap in the 2nd period.

Fig.5 shows strong nonlinearity and nonuniformity of plotted surface. Notably, such oscillations are typical not for densities only, but for all other mentioned properties as well. In principle, if we will switch to long-period table, then these oscillations will become smoother, but such switching is incorrect, because it will require choice of

elements from 4th and higher periods for formation of the "embryo", what looks unobvious. Construction of table from more reasonable "embryo" elements of 2nd and 3rd periods will in any case lead to short-period type of the table.

Moreover, there also exists the problem that cannot be resolved even by transfer to long-period table. Let us consider the fragment of Fig.5 related to 2nd period. Here (Fig.6) it is seen that between starting elements in this period (Li, Be, B, C) and further sequence (starting from N) there is a gap in densities (by a few orders of magnitude – values in vertical axis are given in logarithmic scale). Similar anomaly behaviour is also observed for other physical properties. Gap position may shift from C/N to C/B zone (e.g. for melting and boiling temperatures of oxides), but is always connected with carbon. Existence of such anomaly ("carbon gap") finally leads to impossibility to construct correctly even the 2nd period of the table and to completely absurd construction of following periods.

Therefore, expansion of data array by introduction of additional physics-chemical properties only decreases the quality of systematization process. Of course, it is possible to try using modern elements characteristics (atomic electron configuration, ionization potential, electro negativity etc.), but it looks incorrect, because this information was not available to Dmitry Mendeleev, and, besides, in this case the data array will artificially contain correct solution, thus "killing intrigue" of this study.

Systematization of "Mendeleev's" Set of Elements

Though the systematization of complete set of elements was unsuccessful, in this pert of the study we tried to systemize the elements set that was known to Mendeleev during his work. This set is different from complete one in following:

- inert gases are completely absent (they were discovered later);
- majority of lanthanides and actinides, as well as heavy elements (heavier this bismuth) are absent;
- few elements from middle periods are also undiscovered (Sc, Ga, Ge).

Furthermore, for some elements Mendeleev has doubtful and incorrect data for atomic weights and valences. Due to aforementioned reasons here we used only data on atomic mass and valences. It should be noted that Mendeleev also used this information as basic in table construction process.

Surprisingly, it was found that in this case systematization is much more simple than that for complete set of elements. First of all, it is explained by absence of inert gases, which actually are placed in the table quite illogically (none of them, except xenon, demonstrate 8+ oxygen valence, which is predicted for this group). Additional advantage is absence of lanthanides, because their position in short-period table also looks quite unusual.

In this case we've managed to reproduce the major part of the table, but here as well we met the effect of "wrong" behaviour in triads of transitional elements. For example, cobalt (quite logically) was "trying" to fill the cell of absent gallium due to coincidence of maximum oxygen valence (3+) with relatively low error in prediction of atomic mass. From chemical point of view it is absolutely evident that cobalt is an analogue of iron and nickel, but not aluminum (as gallium), but such "chemical" understanding cannot be described in data array within existing data structure. Anyway, incorrect cobalt positioning led to distortions in further construction of the table. The same may be told about triads of noble metals.

Therefore, we've attempts to modify initial data array, based on exclusion of the most "odious" elements, particularly:

- all elements from transitional triads, except first ones (Fe, Ru, Os), were excluded;
- present lanthanides were excluded (except La and Ce only).

Moreover, for copper and gold the basic oxygen valence 1+ was stated, though their actual maximum valences are higher (2+ for Cu and 3+ for Au). It was done to reveal the fact that Cu and Au are analogues of silver.

The result of systematization in this case is shown in Fig.7.

Period	Group							
							H	
	Li	Be	B	C	N	O	F	
	Na	Mg	Al	Si	P	S	Cl	
	K	Ca		Ti	V	Cr	Mn	Fe
	Cu	Zn			As	Se	Br	
	Rb	Sr	Y	Zr	Nb	Mo		Ru
	Ag	Cd	In	Sn	Sb	Te	I	
	Cs	Ba	La	*Ce*	Ta	W		Os
	Au	Hg	Tl	Pb	Bi	*U*		
				Th				

Fig.7. Result of systematization of "Mendeleev's" set of elements.

It is seen that systematization quality is quite high. Practically all elements are placed in cells, where they should be. Exclusion is made by uranium which actually should be situated in V-7 cell after thorium, but this error is not important (Th and U actually should be placed in separate subgroup of actinides, which is stipulated in this type of the table). Placement of thorium and cerium also does not look formally correct, but actually it is quite usual for them to demonstrate 4+ valences, what gives the ground to position them in the IV-th group of basic table. Such their dual behaviour is well known and is defined by objective specifics of their electronic structure, so such placement may be accepted as appropriate. It is curious, that D.I. Mendeleev the same as also our program, has placed in the initial kind of the table Thorium and Cerium in 4-th group. Moreover, in the same group he has placed and Lantan [2,3]. We shall note that our program has placed Lantan on a correct place in third group.

Prediction of Undiscovered Elements Properties

Special attention should be paid to prediction abilities of the constructed table. As it is seen from Fig.7, after systematization few cells inside the table were left unfilled. These cells strongly correspond to existing elements, that were undiscovered at the time of Mendeleev's study. To predict the properties of missing elements we used the described algorithm. In this case only obtained data values being inside the normalized range ([0,1]) were chosen.

Result of such prediction is quite impressive. First of all, 5 elements that must be positioned inside the table body were clearly shown (their positions in the table are shown at Fig.7 by crossed cells). Description of predicted values is given in Table 1.

Table.1

Position column/string	Atomic mass		Oxygen valence		Hydrogen valence		Real element
	forecast	fact	forecast	fact	forecast	fact	
3/4	43,90	44,95	3	3	-	-	Sc
4/5	69,00	72,59	4	4	4	4	Ge
3/5	65,20	69,72	3	3	-	-	Ga
7/6	101,10	98,91	7	7	1	-	Tc
7/8	176,80	186,20	7	7	-	-	Re

It is seen that coincidence between predicted and actual property values is quite good. Moreover, during analysis of predictions that were excluded from consideration, because the predicted values were found to be outside normalized range, we selected the group of similar elements, which formally should have been positioned in 8th group and have formal valences 8+ for oxygen and 0 for hydrogen (shown by shadowed cells at Fig.7). The reason of their exclusion was zero hydrogen valences, what was considered as inappropriate property value. Actually these predictions are strongly equivalent to the group of inert gases and good coincidence between predicted and actual properties is seen here as well (see Table.2). Furthermore, "wrong" prediction of zero hydrogen valences in this case achieves real physical sense – inert gases do not form hydrogen compounds in reality.

Table.2

Position column/string	Atomic mass		Oxygen valence		Real element
	forecast	fact	forecast	fact	
8/1	6,00	4,00	8	-	He
8/2	20,05	20,18	8	-	Ne
8/3	35,80	39,95	8	-	Ar
8/5	80,30	83,80	8	2	Kr
8/7	138,80	131,30	8	8	Xe

It is interesting that after selection of predictions no forecasts were made non-existent elements, i.e. the algorithm has made no attempts to fill empty cells of 1st period and cells to the right and to the left of the table body.

Conclusion

In general we may state that proposed algorithm BIDIMS (under definite assumptions and modifications) successfully managed to systemize "Mendeleev's" set of elements and, in fact, repeated the discovery of Periodic Law in a form, which was possible in Mendeleev's work period. Performed study, nevertheless, is not diminishing Mendeleev's achievements in any extent. First of all, used assumptions and modifications were based on intuitive and forced decisions, having no formally strong grounds. In second, the basic decisive properties (atomic mass, valences) were chosen the same as ones used by Mendeleev. And the most important – the essence of genius Mendeleev's discovery is proposition, that existing element may be systemized in form of two-dimensional table. We used this proposition as acknowledged fact in our study.

Acknowledgments

Russian Fund of Basic Researches supported the work, grant № 02-01-00082

Bibliography

1. A.N. Zagoruiko and N.G. Zagoruiko. Experiments on Rediscovering of Mendeleev's Low by Using Computer // Strukturnyi Analiz Simvol'nikh Posledovatel'nostei (Vychislitel'nye Sistemy 101, Ak. Nauk SSSR, Novosibirsk, 1984), p. 82-90. (In Russian).
2. A.A. Makarenia. D.I. Mendeleev and physicochemical sciences. Atomizdat. Moscow, 1972, 256 p. (In Russian).
3. http://chemlab.pc.maricopa.edu/periodic/foldedtable.html
4. N.G. Zagoruiko. Applied Methods of Data and Knowledge Mining. Ed. Inst. of Mathematics. Novosibirsk, 1999. 268 p. (In Russian).

Authors' Information

Andrey N. Zagoruiko – Institute of Catalysis of SD RAS, Russia, 630090, Novosiborsk, pr. Lavrentieva, 5

e-mail: zagor@catalysis.nsk.su

Nikolay G. Zagoruiko – Institute of Mathematics SD RAS, Russia, 630090, Novosibirsk, pr. Koptyug, 4

e-mail: zag@math.nsc.ru