RHA(T)-System for Coding of Discrete Distributions and Their Alteration Processes

Tomas G. PETROV,
Faculty of Geology, Saint Petersburg University, Saint Petersburg, Russia

and

Sergey V. MOSHKIN
Open Company "Infocomm" firm, Saint Petersburg, Russia

ABSTRACT

Brief description of RHA(T) information language – system designed for discrete distributions coding, analysis compositions of any nature, their ordering and description of their alteration processes is given. Here, $R$ is rank formula of composition – a sequence of components by reduction of their value (contents); $H$ is Shannon information entropy – complexity measure; $A$ – anentropy – measure of purity; $T$ – tolerance – measure of ultrapurity. The system was tested on objects, the components of which were chemical elements, molecules, minerals, population ages, nationalities etc. Lexicographic – after special alphabets – ordering of composition descriptions enables a linear hierarchic, periodic ordering of rank formulas. RHA(T) system in the chemical version of its utilization, being a System for Compositions representation, is a development of the Periodic System of Elements included by it as an alphabet. RHA(T) characteristics are used as the most adequate ones for studying mixing and separation processes constantly taking place in nature and in engineering.

Keywords: coding, discrete distribution, entropy, anentropy, complexity, purity, mixing, separation.

The system RHA(T) (RHA in the initial version) [2–3], removes a great number of difficulties in these areas, being a general-purpose method for conceptual coding of compositions of any nature, creation of unambiguously structured databases, as well as using a diagram for integral mapping of fundamental processes of composition changes; mixing and separation [4].

2. RANK FORMULA

Rank formula $R$, a semi-quantitative characteristic of composition, is a sequence of event symbols, or elements in composition of an object, by decrease of their occurrence frequency (concentration) $p_i$. In case of contents equality, equal sign is put between them, and components are arranged according to the accepted alphabet. For chemical compositions, alphabet is represented by the sequence of elements in the Periodic System of Elements. Rank formula is written in a row. Rank formulas are situated in column under each other. $R$ length denoted by $n$ is taken for a detail measure of composition study and serves as $R$ length standard when choosing types of events for further calculations. This is crucial, as it is often required to compare analyses of different lengths, with different component lists and with large differences in the content values. Thus, chemical analyses may include the content values differing by several of exponent. $R$ is a name of sector in the composition diagram – simplex divided by median or hyperplanes median (Fig. 1).

![Fig.1. Rank formulas of carbon-tungsten-iron alloys as sector names of a triangle diagram partitioned by medians.](image-url)
Ordering of $R$'s assemblies is made using special alphabets. In this case, $R$ is taken as a "word", where element (component) symbols are "letters". The Periodic System of Elements as the best fixed sequence of elements is taken for a chemical alphabet. Correspondingly, "H", "He", "Li" are the first three letters of this alphabet. For interval composition number axis is an alphabet. Vertically ordered alphabetic sequence of such "words" generates a linear hierarchic, periodic classification of composition rank formulas (Tab. 1).

Hierarchic character of classification is determined by two factors. The first one is that each rank formula with $n$ detail belongs to a single rank formula with $n-1$ detail. The second one is that the hierarchic character is shown in the rank formula itself. Thus, the first rank ($R_1$) is occupied by the most widespread element in the given system, for example, oxygen O. Class of the second rank $R_2$ – OHa – silica-oxygenic enters into it. In its turn, $R_3$ class – OSiAl – alumino-silica-oxygenic enters into the latter class. System periodicity is shown in large enough assemblies in group arrangement of similar objects. In Tab.1, the system periodicity is exemplified by carbonate arrangement.

From combinatorial analysis positions rank formulas are distribution from $N$ alphabet symbols on $n$. Correspondingly, for example, for chemical compounds with $n-1$, $N$ equals 38. With $n$ increasing, the number of rank formulas rises sharply and for $n=10$ their number increases to $7.7*10^{19}$. The actual number should be much smaller for cosmochemical and technological reasons. Generally, such a large number of possible $R$’s also does not interfere with their use, as a similar number of words that can be generated combinatorially in natural languages does not interfere as well.

The text of $R$-dictionary [5] was ordered according to the described above alphabet - in the collection of information about the theoretical chemical compositions of all known minerals. The collection included more than 4,500 records. $R$-dictionary, as far as we know, is the first publication in which search characteristic is represented by strictly uniformly arranged data on the chemical composition of minerals. This method allows, after the available chemical formula (not only of minerals) is transformed into a rank formula, determination of the mineral name in 92% of cases. The rest 8% are polymorphs – minerals with the same chemical composition but different structure. Optical or X-ray methods are necessary for their determination. According to the $R$-dictionary materials, quantities of minerals with the given length of rank formulas were calculated (Fig. 2).

![Fig. 2. Distribution of minerals after the rank (chemical) formulas length](image)

The upper curve - distribution of all minerals; the lower one, only "new" minerals discovered in recent years. When preserving the rules of new minerals distinguishing, curve shape should not change significantly. Therefore, we can, first, say that in conditions of accessible for studying part of the Earth's crust, maximum number of minerals forms as chemical compounds containing 4-5 chemical elements. We can surely say that statistics of minerals, for example moon ones, will be different. Second, a weak minimum on the upper curve and distinct one on the lower curve indicates that this phenomenon is clearly not a matter of nature. Apparently, this fact is an evidence of the initiated exhaustion of the minerals diversity for discovery of the new ones.

Rocks can be regarded as mixtures of minerals with different compositions. Problems of their unequivocal description and classification have long existed and have been solved separately for igneous, sedimentary, metasomatic, metamorphic rocks. Using sequences of mineral symbols after reduction of contents in rock as rank formulas, and $R$-dictionary as an alphabet, one can construct a common systematic of any rocks, for which mineral-molecular compositions exist (or can be estimated). Accordingly, chemical-molecular information on the technology, food, pharmaceutical products may be organized in such a way.

Rank formula is the first step for composition description. The second is introduction of quantitative characteristics.

**3. QUANTITATIVE CHARACTERISTICS OF DISTRIBUTIONS**

The 1st quantitative characteristic of composition with $n$ detail is C. Shannon information entropy as a composition complexity measure, $H = -\sum p_i \log p_i$. Information entropy is an analogue of thermodynamic entropy of mixing [6]. Standardized to the interval 0-1 entropy is calculated by formula $H_i = \frac{H}{H_{\text{max}}}$. Isolines of standardized entropy in the three-component diagram are shown in Fig. 2.

![Fig. 2. Entropy isolines of the three-component diagram](image)

When the system is split into two with composition change, entropy of at least one of the final products, as it was proved [7], is lower than the entropy of the original system. When two different compositions are mixed, entropy of the resulting composition is higher than that of at least one of the original ones. Conclusions of these theorems: during separation, complexity significantly decreases statistically; during mixing, composition complexity significantly increases statistically.

Entropy provides for a preferential consideration of large and medium in content components and is low sensitive to the small
ones. Therefore, compositions differing strongly, especially in small components, can have the same \( H (En) \) value (Fig. 3). In addition, compositions of objects with the same \( R \) and \( En \) but relating to different types exist. In particular, this is the situation with some shales (rocks) and muscovites (mineral). This fact has led to the introduction of the second characteristics of ranked distribution of component compositions.

Fig. 3. Compositions variants at \( En = 0.376 \). 1 – theoretical composition – the first component 82%, the rest 2% each; 2 and 3 – rocks; 4 – mineral of “simple” composition; 5 – theoretical composition – the first two components 47% each, third 5.96, and the rest 0.005% each.

For preferential estimation of small components smallness, the purity measure of compositions, anentropy \( A \) is introduced \([1–4]\). In the simplest case, \( A = - (\Sigma p_i \lg p_i)/n – \lg n \). Anentropy values belong to the range 0 ÷ +\( \infty \). Minimum value \( A = 0 \) is obtained when \( p_1 = p_2 = p_3 = \ldots = p_n = 1/n \). Maximum \( A = +\infty \) is get if in the analysis at least one \( p_i \neq 0 \). For anentropy reduction to the interval 0÷1, anentropy is divided by \( A \) of the “analytically ideally pure system”. Such composition consisting of \( n-1 \) admixtures in contents equaling half sensitivity of a certain standard investigation method, and one component completing their sum to 1 is taken for such system. For chemical analyses of rocks and minerals, half sensitivity is taken for \( p = 0.00005 \). We get \( A_n \) independent of logarithm base. The name “anentropy” reflects its similarity to entropy and the opposite with respect to small components.

Fig. 4. shows a three-component diagram with plotted anentropy isolines. It is seen that \( A_n \) varies slightly at compositions commensurability and increases sharply when approaching the composition, in which one component dominates. In chemistry, it is an “elementary substance”. Let us note that the most effective methods for improving purity of chemical elements leave in the matter dozens of elements – admixtures analyzed by modern methods.

Comparability of \( A_n \) values for analyses with different \( n \) is provided for by \( R \) length standardization. For identification of the overwhelming majority of geological objects and for solving of many genetic problems, \( n = 10 \) detail was accepted. This corresponds to consideration of 99.5-99.9% of matter records in the system.

For describing compositions, in which taking into account of extra small component contents is necessary, one more characteristic – measure of ultrapurity – “tolerance” \( T \) was introduced \([8]\). It is calculated from the formula \( (\Sigma (1/p_i))/n \). In case of weak discrimination of compositions in \( En-A_n \) diagram, one passes to \( En-T \) diagram.

All three characteristics \( at \ the \ given \ composition \) are independent of the phase state, temperature, pressure, and structure of an object; therefore, they allow reflecting processes of compositions change in systems of any nature. At the same time, values, especially \( A_n \) and \( T \) for the systems resulting from separation, the mechanism of which is generally selection, depend on the speed of this process. The higher its speed, the less perfect separation, the more difficult and less pure the composition becomes, the higher \( En \) values and the lower \( A_n \) values and \( T \), approaching the original composition.

Investments of individual concentration into \( H, A \) and \( T \) are shown in Fig.5.

As one can see, maximum contributions to the integral characteristic is made by \( 1/p \). That is, the use of tolerance will monitor the operational process when obtaining high purity matters. In other cases – to capture the effect of rare events on the overall behaviour of the system.
4. EnAn (or EnT) DIAGRAMS

EnAn (or EnT) diagrams are used:

1) for reflection of disordered datasets, features of point distinguished in this case, heterogeneities of the distribution density of points, the nature of arcs direction are revealed, manifestation of anomalous analyses;

2) for ordered ones, particularly, for processes reflection. En increase and An decrease correspond to mixing processes (mixture preparation: pharmaceutical, food etc.; exhaust gas emission into the atmosphere, river inflow into seas…). H decrease and A increase are peculiar to separation processes (any concentration, purification, crystallization etc.). Several typical trajectories of entropy characteristic changes under different processes of compositional changes are shown in Fig. 6.

Processes of minerals and rocks evolution shown here are crystallization processes, which are typical separation processes. They take place under the increment of content of the first components in R’s and decrease of the last ones. Direction of trajectory of the national composition change in Turkmenistan is the same. Mixing processes are exemplified by the national composition change in Estonia and water dissolution of potassium nitrate.

Fig. 6. Evolution of compositions, (calculations for 10 elements) 1 – water, 2,3,4,5 – water dissolubility of potassium nitrate at temperatures 0, 20, 40, 60°C; 6 – potassium nitrate (KNO₃); 7 – mineral scheelite (CaWO₄) change in composition from the crystal center to the periphery [9]; 6 – magmatic rocks from gabbro to granite [10]; (calculations for 4 components): national composition [11]; 7 – Turkmenistan (1959-1979), 6 – Estonia (1959-1979).

Such diagrams enable comparison of the processes taking place in different systems. As we see, evolution of compositions of individual minerals in rocks and composition of rocks hosting them etc. may be traced.

5. RHAT IN GENERAL

RHAT assembly may be considered a “soft” system of coordinates specifying location of the composition area in the compositions diagram (in contrast to “hard” systems of coordinates describing point position).

From positions of semiotics RHAT it alphanumeric-sign symbol system of the description of compositions and their alteration.

From stand point of information theory this method is rank-entropy characterization of system composition on subset their dominant components.

Entropy used in dozens of sciences, together with anentropy mathematically related with it, enabled construction of a diagram, which can be used to depict the processes of composition change without restrictions on the nature of changing systems.

The main advantage of using RHAT consists in simplicity and unity of the representation way and ordering the compositions of objects of any nature, which allows organizing quantitative analytical data as hierarchical, periodic systems. At the same time, systematizing of objects is performed regardless of whether they have names. The method was mainly developed on compositions of geological objects, where over twenty alternatives of its utilization are implemented [12]. Moreover, its potential is tested on age and national distributions of population, in linguistics, as well as on changes of personal income distributions, as a two-dimensional alternative to the one-dimensional Gini coefficient.

RHAT-method for chemistry is a development of the Periodic System of Elements. The method provided the possibility of a uniform ordering of all available (and pending) information on the chemical compositions of objects of any nature. Organization of the Hierarchical Periodic System of Compositions is correlated to the structure and information on the composition embedded into the Periodic System of Elements after 17 entries [13]. Main similarities and differences are shown.

For any chemical composition, there is a strictly defined place in the System of Compositions, as in a dictionary or encyclopedia. Properties of the Periodic System of Elements are inherited in the coding system: average atomic weights of compositions grow from the beginning to the end of the System of Composition.

RHAT system and PETROS was used for creation of a databank «Chemistry of Nature Objects». It includes chemical compositions of minerals, rocks, ores, waters, gases, meteorites, and other space bodies (46,000) records. Databank sampling is represented in Table.

Low En and high An values of the Sun, water, and minerals (pyrite) bear the evidence of strong differences of the first components in R from the last ones. High En and low An values correspond to a more uniform distribution of component contents (carbonatite, kimberlite, bone fishes).

In the Internet, there is an integrated catalogue of rocks and rock-forming minerals (4,300), minerals (micas – 1,600, pyroxenes – 900, tourmalines 800 etc.), as well as a catalogue of mineral compositions of alkalic rocks combined with the catalogue of chemical compositions (480 records).
Work on the method is provided by Petros 3 software package, the recent version of which allows using alphabets 100 in 100 symbols each [14].

Brief description of RHAT method, collections of compositions of different minerals and rocks as well as their codes in RHAT form are available at the web-site: http://geology.spbu.ru/department/scientific/rha-language-method

ACKNOWLEDGMENT

The authors would like to thank N.I. Krasnova, I.S. Sedova, A.G. Bulakh, V.V. Gordienko and V.N. Dech for active assistance in the method development, as well as to thank all, whose active antagonism contributed to searching of the optimal substantiation and description.

Table. Data sampling from Databank «Chemistry of Nature Objects»

<table>
<thead>
<tr>
<th>Rank</th>
<th>Formula</th>
<th>En</th>
<th>An</th>
<th>Object</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Na2CO3</td>
<td>3</td>
<td>1</td>
<td>Sodium carbonate</td>
</tr>
<tr>
<td>2</td>
<td>Na2SO4</td>
<td>2</td>
<td>1</td>
<td>Sodium sulfate</td>
</tr>
<tr>
<td>3</td>
<td>CaCO3</td>
<td>4</td>
<td>1</td>
<td>Calcium carbonate</td>
</tr>
</tbody>
</table>

REFERENCES


