INFORMATION-ANALYTICAL SYSTEM FOR DESIGN OF NEW INORGANIC COMPOUNDS

Nadezhda Kiselyova, Andrey Stolyarenko, Vladimir Ryazanov, Vadim Podbel'skii

Abstract: The principles of design of information-analytical system (IAS) intended for design of new inorganic compounds are considered. IAS includes the integrated system of databases on properties of inorganic substances and materials, the system of the programs of pattern recognition, the knowledge base and managing program. IAS allows a prediction of inorganic compounds not yet synthesized and estimation of their some properties.

Keywords: information-analytical system, knowledge discovery in databases, design of new inorganic compounds, pattern recognition, computer learning, knowledge base, databases on properties of inorganic substances and materials.

ACM Classification Keywords: J.6 Computer-aided Design, J.2 Computer Applications & Chemistry, H.2.8 Scientific Databases, I.2.6 Analogies, I.2.6 Learning, I.2.4 Knowledge Representation Formalisms and Methods, C.2.5 Internet.

Introduction

The problem of prediction of formation of new compounds and calculation of their properties is one of the most important tasks of inorganic chemistry. Any successful attempt of design of compounds not yet synthesized is of the large theoretical and practical importance. The problem of design of new inorganic compounds can be formulated as follows: it is necessary to find a combination of chemical elements and their ratio (that is, qualitative and quantitative composition) for making (under the given conditions) the predefined space molecular or crystal structure of compound allowing a realization of necessary functional properties. Only properties of chemical elements and data about other already investigated compounds should be used as initial information for calculations. Thus, the problem is concerned with a search for regularities between properties of chemical systems (for example, properties of compounds) and properties of elements, which form these systems.

The decision of a task of design of new inorganic compounds presents severe difficulties. The main difficulty is an extreme complexity of dependences relating property of inorganic compounds with properties of chemical elements. The traditional way of the decision of this task is associated with quantum-mechanical methods, which are based on the Schrodinger's equation. However in most cases the accurate solution (in analytical functions) of the latter for certain inorganic substances is fraught with great mathematical difficulties, which were been overcome only for the simplest systems. Therefore various approximated methods, as a rule, are used. These methods very much frequently do not give desirable results.

On the other hand, the chemistry had accumulated large information on properties of inorganic substances. There are periodic regularities between properties of compounds and properties of elements, which are included into their composition. This supposition is a consequence of the Periodic Law. Moreover, it is obviously, that already known compounds should be in accordance with these periodic regularities. The aims of our researches are development of methods and creation of computer system for search for these periodic regularities on the basis of analysis of information about already known substances accumulated in databases on properties of inorganic substances and materials. The found regularities are used for design of new inorganic compounds – analogues of already synthesized substances.

Selection of Methods of Search for Regularities in Information of Databases on Properties of Inorganic Substances and Materials

The methods of computer learning in pattern recognition are one of the most effective means of search for regularities in the large arrays of the chemical data [Kiselyova, 2005; Savitski and Gribulya, 1985]. In this case it is possible to connect some discrete parameters of inorganic compounds (for example, possibility of formation of compound or type of its crystal structure under normal conditions) with properties of elements, which are included into their composition, and also to get a threshold estimation of some numerical properties (for example, estimation of the melting point of compound at atmospheric pressure - above or below than certain threshold). It is important, that the fulfillment (though also not so strict) of basic hypothesis of methods of pattern recognition - hypothesis of compactness - is a consequence of the Periodic Law. Let an each compound corresponds to a point in multi-dimensional space of properties of elements. Owing to periodicity of properties of chemical elements points, which correspond to combinations of close on properties elements, combining into compounds, form compact clusters. Thus, the task of search for regularities connecting property of inorganic compounds with properties of chemical elements can be reduced to a problem of computer learning in pattern recognition. In this case the analysis of the information about already known compounds, which are represented as a set of values of properties of chemical elements, allows discovery of classifying regularities. The latter allow separation of known compounds into predetermined classes. It is possible to predict new compounds and estimate their unknown parameters by substitution of the property values of the appropriate chemical elements into the found regularities.

The principal problems at application of methods of pattern recognition to the decision of tasks of inorganic chemistry are following:

1. Small informativeness of attributes - properties of chemical elements.
2. The strong correlation of these attributes owing to their dependence on common parameter - atomic number of chemical elements (it follows from the Periodic Law).
3. Omissions in values of attributes.
4. In many cases - the large asymmetry of a size of classes of training set.
5. Sometimes feature description includes non-numerical attributes.
6. Possibility of experimental mistakes of classification in training sets.

In connection with the above-stated peculiarities of subject domain the search for methods and algorithms of pattern recognition allowing correct solution of these problems was one of the basic tasks of development of information-analytical system (IAS) for computer-aided design of inorganic compounds. It was established during testing various algorithms of computer learning for concrete tasks that it is impossible to specify beforehand, what algorithm is most effective at the decision of the certain chemical task of design of inorganic compounds. Quite often programs, which well have classified training set, obtained bad results at the prediction of unknown compounds. In this connection the most effective way of decision of tasks of predicting properties of new inorganic compounds is concerned with methods of recognition by collectives of algorithms [Zhuravlev et al., 2006]. At synthesis of the collective decision it is possible to compensate mistakes of separate algorithms by the correct predictions of other algorithms. Hence, the developed information-analytical system includes a set of the programs realized algorithms of various types, and also different strategies of collective decisions making.

Other way of increase of accuracy of predicting is a use of dependence of properties of chemical elements on atomic number. On the one hand, this fact complicates a task of search for separate properties that are the most important for classification of because of strong correlation of all used parameters of elements forming feature description. On the other hand, the classifying regularities including values of any subset of properties of chemical elements, which are used for the description of inorganic compounds, should in principle give identical results of classification. I.e. the results of the prediction with use of various subsets of properties of elements should, basically, coincide. This fact allows an additional possibility of collective decision making but already on the basis of collective of feature descriptions which was obtained as a result of division of initial set of properties of chemical elements on partially crossed subsets.

The problem of filling omissions also is partially solved with use of periodic dependences of parameters of elements. Replacement of the omission by average value of given parameter for two chemical elements that are nearest (within the range of group of Periodic System) to the element with omission is used.

After testing the programs the following software of pattern recognition were included into information-analytical system:

- a wide class of algorithms of system RECOGNITION developed by A.A.Dorodnicyn Computer Center of Russian Academy of Sciences (CCAS) [Zhuravlev et al., 2006]. This multifunctional system of pattern recognition includes the well-known methods of *k*–nearest neighbors, Fisher's linear discriminant, linear machine, multi-level perceptron (neural networks), support vector machine, genetic algorithm, and the special algorithms which were developed by CCAS: estimates calculation algorithms, LoReg (Logical Regularities), deadlock test algorithm, statistical weighted syndromes, etc.

- system of concept formation ConFor developed by V.M.Glushkov Institute of Cybernetics of National Academy of Sciences of Ukraine [Gladun, 1995, 2000, 2005]. The system is based on special data structure in a computer memory named as growing pyramidal networks.

It is important, that system RECOGNITION [Zhuravlev et al., 2006] is equipped with a set of algorithms of the decision of tasks of recognition by collectives of various algorithms. In this case task of recognition is decided in two stages. At first various algorithms, which are included into system, are applied independently. Further an optimum collective decision is made automatically with the help of special methods - "correctors". Some of methods of synthesis of the collective decisions - Bayesian corrector, convex stabilizer, some heuristic methods, etc. are used as correctors.

Databases and Knowledge Base of Information-Analytical Systems

The information basis of IAS (fig.1) is the integrated system of databases on properties of inorganic substances and materials [Dudarev et al., 2006; Kiselyova et al., 2005], which now includes:

- DBs containing the brief information on the most widespread properties of inorganic compounds and chemical elements:

1). DB on properties of inorganic compounds "Phases" [Kiselyova, 2005; Kiselyova et al., 2006] which now contains the information on properties more than 43, 000 ternary compounds (i.e. compounds formed by three chemical elements) and more than 15, 000 quaternary compounds, that was extracted from about 20, 000 publications.

2). DB on properties of chemical elements "Elements" which includes the data on more than 90 parameters.

- Specialized DBs which contain the detailed information on industrially vital substances and materials:

1). DB of phase diagrams of systems with intermediate semiconducting phases “Diagram” [Kiselyova, 2005; Khristoforov et al., 2001], that contains information on the most important pressure-temperature-concentration phase diagrams of semiconducting systems evaluated by qualified experts and also on the physical-chemical properties of the intermediate phases. Now DB contains the detailed information on several tens binary and ternary systems extracted from 2000 publications.

2). DB on substances with significant acousto-optical, electro-optical and nonlinear-optical properties “Crystal” [Kiselyova, 2005; Kiselyova et al., 2004] which now includes the information on parameters more than 100 materials.

3). DB on width of the forbidden zone of inorganic substances “Bandgap” [Dudarev et al., 2006] which now contains the data on more than 2, 000 substances.

Cumulative volume of DBs is ~7 GB. All these databases are accessible from Internet (www.imet-db.ru).

The knowledge base (KB) of information-analytical system includes the relational tables containing regularities found during computer learning with the indication about common chemical composition of compounds, set of attributes included into regularity, parameter to be predicted, used algorithm, and also service information (date of updating, surname of the expert who carried out computer learning, etc.). Both the integrated system of DBs on properties of inorganic substances and materials and knowledge base are realized with use of DBMS MS SQL Server. DBs, which are incorporated into the integrated information system, use various DBMS [Dudarev et al., 2006].

The information-analytical system for design of inorganic compounds is intended for users of two levels. Firstly it is a reference system for ordinary specialist. Secondly IAS is a tool for expert estimating the chemical information for computer learning and carrying out a search for regularities in data (fig.1). In last case owing to use of knowledge and experience of the highly skilled experts the mentioned above problems of selection of the most important attributes for the description of compounds and filtration of mistakes of classification of objects of training set are partially decided.

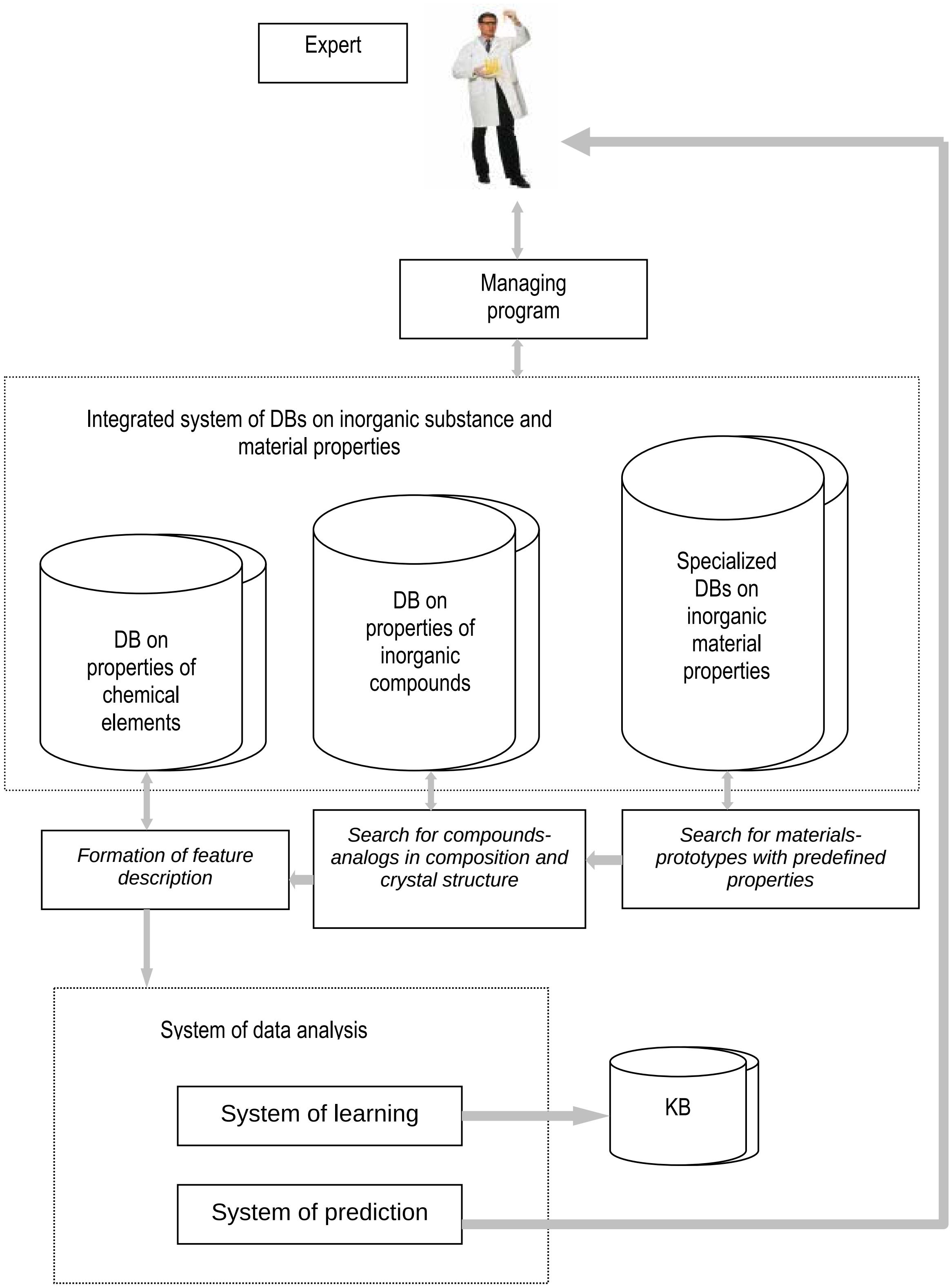
Expert
Managing program
Integrated system of DBs on inorganic substance and material properties
DB on properties of chemical elements
DB on properties of inorganic compounds
Specialized DBs on inorganic material properties
Formation of feature description
Search for compounds-analogs in composition and crystal structure
Search for materials-prototypes with predefined properties
System of data analysis
System of learning
System of prediction
KB

Fig.1. Principal schema of IAS

Program Realization of Information-Analytical System

Feature of program realization of IAS is the design of client module on the basis of the Web-interface completely. The users work with IAS using only Web-browser. Thus, the users do not need to install of any additional programs. It also facilitates an expansion of system by new methods and functionalities: the changes are done only in server where the system is located. Interaction between predicting subsystems, which realize various methods of learning and predicting, and subsystem of the graphic Web-interface is realized on the basis of the interface module-broker or the "shell" giving all necessary functions of system, that also corresponds to the ideas of SOA-approach. The operation of processes of training and recognition is realized by means of asynchronous Web-services. At design of IAS it is necessary to provide storage of the information about the programs of data analysis and methods of recognition realized in them. These data are necessary for a correct invocation of functions of programs and setting of learning methods. The relational DB called as metabase was used for data storage. The knowledge base is realized using SQL-server and Web-server. Web-server is intended for storing all necessary files using special format for subsequent their application to recognizing. SQL-server stores complete information on the obtained regularities.

Conclusion

The information-analytical system, created by us, allows solution of two important tasks of inorganic chemistry. First, it allows the partially automation of analysis of the huge experimental information, accumulated by chemistry, for search for regularities in the data and subsequent design of new compounds with predefined properties. Secondly, it expands opportunities of traditional DBs on properties of substances and materials, giving the user not only information on the already investigated substances, but also predictions of some substances not yet synthesized and estimation of their properties. Essential advantage of developed IAS is Internet-access. In this case user receives operative access to "alive" data and regularities. With the help of IAS it was possible to predict some new inorganic compounds and to estimate their some properties.

Acknowledgements

Financial support from RFBR (grants №06-07-89120 and 05-03-39009) is gratefully acknowledged. We should like to thank our colleagues Prof. V.P.Gladun, Dr. V.Yu.Velichko of the Institute of Cybernetics of the National Academy of Sciences of Ukraine, Prof. V.S.Zemskov, Dr. V.A.Dudarev, Dr. I.V.Prokoshev, V.V.Khorbenko of the Institute of Metallurgy and Materials Science of Russian Academy of Sciences, and Dr. O.V.Sen'ko of the Computer Center of the Russian Academy of Sciences for their help and support.

Bibliography

[Dudarev et al., 2006] V.A.Dudarev, N.N.Kiselyova, V.S.Zemskov. Integrated system of databases on properties of materials for electronics. Perspektivnye Materialy, 2006, N.5 (Russ.).

[Gladun, 1995] V.P.Gladun. Processes of Formation of New Knowledge. SD "Pedagog 6", Sofia, 1995 (Russ.).

[Gladun, 2000] V.P.Gladun. Partnership with Computer. Port-Royal. Kiev, 2000 (Russ.).

[Gladun, 2004] V.P.Gladun. Growing pyramidal networks. Novosti Iskusstvennogo Intellekta, 2004, №1 (Russ.).

[Khristoforov et al., 2001] Yu.I.Khristoforov, V.V.Khorbenko, N.N.Kiselyova, et al. Internet-accessible database on phase diagrams of semiconductor systems. Izvestiya VUZov. Materialy Elektron.Tekhniki, 2001, №4 (Russ.).

[Kiselyova, 2005] N.N.Kiselyova. Computer Design of Inorganic Compounds. Application of Databases and Artificial Intelligence. Nauka, Moscow, 2005 (Russ.).

[Kiselyova et al., 2005] N.N.Kiselyova, V.A.Dudarev, I.V.Prokoshev, et al. The distributed system of databases on properties of inorganic substances and materials. Int.J."Information Theories & Applications", 2005, v.12.

[Kiselyova et al., 2006] N.Kiselyova, D.Murat, A.Stolyarenko, et al. Database on ternary inorganic compound properties "Phases" in Internet. Informazionnye Resursy Rossii, 2006, N.4 (Russ.).

[Kiselyova et al., 2004] N.N.Kiselyova, I.V.Prokoshev, V.A.Dudarev, et al. Internet-accessible electronic materials database system. Inorganic Materials, 2004, v.42, №3.

[Savitski and Gribulya, 1985] E.M.Savitski and V.B.Gribulya. Application of Computer Techniques in the Prediction of Inorganic Compounds. Oxonian Press Pvt.Ltd., New Delhi-Calcutta, 1985.

[Zhuravlev et al., 2006] Yu.I.Zhuravlev, V.V.Ryazanov, O.V.Senko. RECOGNITION. Mathematical methods. Software System. Practical Solutions. Phasis, Moscow, 2006 (Russ.).

Authors' Information

Nadezhda Kiselyova – A.A.Baikov Institute of Metallurgy and Materials Science of Russian Academy of Sciences, P.O.Box: 119991 GSP-1, 49, Leninskii Prospect, Moscow, Russia, e-mail: kis@ultra.imet.ac.ru

Andrey Stolyarenko –- A.A.Baikov Institute of Metallurgy and Materials Science of Russian Academy of Sciences, P.O.Box: 119991 GSP-1, 49, Leninskii Prospect, Moscow, Russia, e-mail: stol-drew@yandex.ru

Vladimir Ryazanov – A.A.Dorodnicyn Computer Center of Russian Academy of Sciences, e-mail: riazanov@ccas.ru

Vadim Podbel'skii – Moscow Institute of Electronics and Mathematics (Technical University), P.O.Box: 109028, B.Trehsvjatitelsky per. 3/12, Moscow, Russia, e-mail: vvp@mitme.ru, vpodbelskiy@hse.ru

AN IDEA OF A COMPUTER KNOWLEDGE BANK ON MEDICAL DIAGNOSTICS

Mery Chernyakhovskaya, Alexander Kleschev, Filip Moskalenko

Abstract: The paper is a description of information and software content of a computer knowledge bank on medical diagnostics. The classes of its users and the tasks which they can solve are described. The information content of the bank contains three ontologies: an ontology of observations in the field of medical diagnostics, an ontology of knowledge base (diseases) in medical diagnostics and an ontology of case records, and also it contains three classes of information resources for every division of medicine – observation bases, knowledge bases, and data bases (with data about patients), that correspond to these ontologies. Software content consists of editors for information of different kinds (ontologies, bases of observations, knowledge and data), and also of a program which performs medical diagnostics.

Keywords: Medical Diagnostics, ontology model, parallel computing, knowledge bank.

ACM Classification Keywords: I.2.1 Applications and Expert Systems, J.3 Life and Medical Sciences.

Introduction

Computer systems for medical diagnostics are one of applications of AI systems. They can help doctors to improve the quality of their work. The task of such systems is to recognize diseases (one or several), with which a patient is ill, basing on the results of patient's observations. The important components of such systems are a confidence subsystem which can show to the doctors the knowledge base of the system and an explanation subsystem which can show to the doctors the information and reasoning way which were used to produce the result.

Two classes of the systems for solving the task of medical diagnostics have been developed by now, which differ by methods that lie in their base. The systems of the first class are based on statistical and other mathematical models – their bases are mathematical algorithms that perform the search of usually a partial correspondence between the symptoms of the current patient and the symptoms of previous patients for whom the diagnoses are known [1 – 4]. However such systems lack the confidence and explanation subsystems.

The systems of the second class are based on expert knowledge. Their algorithms operate with the information about the patient and with the knowledge about diseases which are represented in a form that is more or less close to the concepts of doctors (and described by expert-doctors). That is achieved by an explicit or implicit usage of ontologies of medical diagnostics. In these systems it is possible to create the subsystems of confidence and explanation which is capable of giving a doctor the results of analysis of patient's state that led to the derived result.