

Relevance Evaluation of Information Retrieval in the Integration of Information Systems on Inorganic Substances Properties

Victor Dudarev^{1[0000-0002-3583-0704]}, Nadezhda Kiselyova^{2[0000-0001-7243-9096]} and Igor Temkin³

¹ National Research University Higher School of Economics, Moscow, 109028, Russia

² A.A. Baikov Institute of Metallurgy and Materials Science of RAS (IMET RAS), Moscow, 119334, Russia

³ National University of Science and Technology MISIS (Moscow Institute of Steel and Alloys), Moscow, 119049, Russia

vdudarev@hse.ru

Abstract. One of the main tasks in the integration of information systems is to provide relevant retrieval of information consolidated from heterogeneous sources. In the field of inorganic chemistry and materials science, set-theoretic methods of searching for relevant information are known. They ensure the construction of a sufficiently high-quality response to user requests. However, the problem of quantifying evaluation of information search relevance in this subject area remains open. This paper proposes an approach to quantifying evaluation of the relevance of information retrieval in integrated systems on inorganic substances and materials properties.

Keywords: relevance evaluation, database integration, inorganic substances.

1 Introduction

The development and use of integrated information systems on substances and materials properties that consolidate information from heterogeneous information sources is worldwide common trend. These systems ensure that specialists are capable to quickly find the required information. When developing such systems, the fundamental thing is data representation method that describes corresponding chemical objects and their properties. Furthermore, chemical objects data representation method, in its turn, determines the class of methods for ensuring the search for relevant information and their functionality. The purpose of this paper is to present a new approach for quantifying evaluation of the relevance of information retrieval for integrated information systems (IS) on inorganic substances and materials properties (ISMP) based on information structures describing the qualitative and/or quantitative substance composition.

Copyright © 2019 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

2 The current state of the problem

2.1 Heterogeneous information systems

The information technologies development and the emergence of powerful hardware and software tools for storing and processing information stimulated works on information systems development in the field of inorganic materials science. As a result, a large number of highly specialized information systems have been developed that are focused on solving problems with due regard for specificity, conditioned by a specific subject domain and research areas of a specific organization developing IS. An example is a number of information systems based on databases developed and maintained by IMET RAS. The IMET RAS information systems core consists of a number of databases which store data on a variety of properties of substances:

- «Diagram» – database (DB) on the phase diagrams of semiconductor systems;
- «Crystal» – DB on the properties of acoustooptical, electro-optical and nonlinear-optical substances;
- «Phases» – DB on the general properties of ternary and quaternary compounds;
- «Bandgap» – DB on the band gap of inorganic substances [1];
- «Elements» – DB on the properties of chemical elements.

These databases are heterogeneous not only by data structures, but also by software and hardware tools ensuring their operation [2]. It should be noted that above mentioned DBs contain extensive information, but in a fairly narrow area. The situation when none of the developed information systems contains a complete set of data on properties of an object (substance or material) and the specialist needs to use several information resources at once to search for the necessary information is typical not only for inorganic materials science, but also for other subject domains.

Obviously, to ensure a high-quality information service for materials scientists, information systems integration in this subject domain is necessary. In Russia, the first successful attempts in this direction were undertaken at the beginning of the century at the IMET RAS for the integration of information systems mostly used by Russian users [3]. The integration allowed a consolidation of information resources for end users and a significant reduction of the time spent by specialists to find the necessary information. The applied consolidation approach was based on the Enterprise Application Integration (EAI) method and showed its efficiency and good scalability when connecting resources developed in different organizations to the integrated information system [8]. For example, «TCS» (Thermal Constants of Substances - reference book on substances thermal constants, developed by the Joint Institute for High Temperatures of Russian Academy of Sciences (JIHT RAS) together with the Moscow State University (MSU)) and «AtomWork» (information system on inorganic substances properties, developed by the National Institute for Materials Science (NIMS), Japan) are among successfully integrated systems [4].

One of the main difficulties in the heterogeneous information system (IS) integration is the diversity of the chemical objects described in them. So, for example, «Diagram»

IS contains information at the level of the chemical system, i.e. a set of chemical elements that form a certain phase diagram of a semiconductor system. Other IS on inorganic substances and materials properties (ISMP) describe the properties at a specific quantitative composition level (with a specific ratio of elements in chemical system), taking into account crystalline modifications of substances, i.e. the quantitative composition of the substance and its crystal lattice are described at this level. Such chemical objects descriptions incompatibility in different IS ISMP dictates the need to use a different description of chemical objects in an integrated IS ISMP, at least it's required to distinguish between several types of chemical objects: chemical systems, substances and their crystal modifications.

2.2 Chemical objects hierarchy

To describe the basic chemical objects of the considered problem domain the set theory is used, taking into account that each subsequent level in the problem domain hierarchy complements the description of the chemical object. The notation is the following: S is the set of chemical systems; C – set of chemical substances, i.e. chemical compounds, solid solutions, heterogeneous mixtures, etc.; M – set of crystal modifications. Then the chemical system is denoted as s (where $s \in S$), the chemical substance is denoted by c (where $c \in C$), and the crystal modifications is m (where $m \in M$).

Having designated second level objects by the «substance» term, we get three-level chemical objects hierarchy: chemical system, chemical substance and chemical modification [5]. As far as information stored in DBs on inorganic substances properties can be considered at chemical system level, for simplicity we'll use this level from the top of the objects hierarchy. So, the chemical objects hierarchy and relationships between chemical objects can be described by means of chemical objects hierarchy in tree form (Fig. 1).

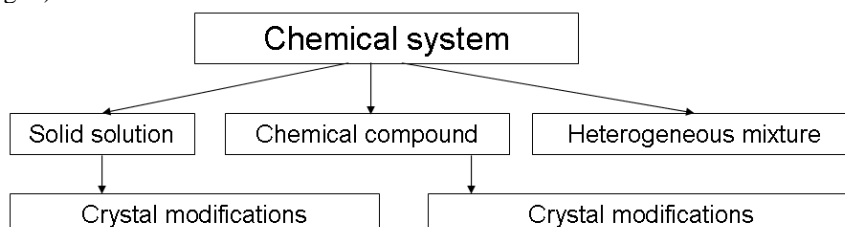


Fig. 1. Chemical objects hierarchy.

Any chemical system s can be represented as a set of chemical elements e_i : $s = \{e_1, e_2, \dots, e_n\}$. Any chemical substance c is defined not only by the set of atoms (chemical elements), but also by their quantitative incorporation into the composition of the compound, solution or mixture. Therefore, any substance c can be represented by a tuple (s, f) , where $s \in S$, and f is a mapping of the set of atoms (chemical elements) that make up the substance, in the set of $\mathbf{R}^* \times \mathbf{R}^*$ pairs that define the minimum and maximum incorporation of a given chemical element in a compound, solution or mixture c .

That is, $f: e_i \rightarrow (\mathbf{R}^*_{min}, \mathbf{R}^*_{max})$, where $\mathbf{R}^* = \mathbf{R}^+ \cup \{x\}$. \mathbf{R}^+ is the set of non-negative real numbers, and \mathbf{R}^* is the set of \mathbf{R}^+ extended by the element x . The element x is used to denote an unknown number, since in the notation of mixtures where the incorporation of components may vary, it is customary to use x to denote an unknown, for example, $\text{Fe}_{1-x}\text{Se}_x$. \mathbf{R}^*_{min} and \mathbf{R}^*_{max} are, respectively, the minimum and maximum concentration of the chemical element e_i in the substance c .

In the case when the concentration of a particular chemical element e_i in the substance c is fixed, then $\mathbf{R}^*_{min} = \mathbf{R}^*_{max}$. Chemical modification m can be represented by a tuple (s, f, mod) , where $s \in S$, $f: e_i \rightarrow (\mathbf{R}^*_{min}, \mathbf{R}^*_{max})$, and mod is the string notation for the crystal modification of a substance – common for integrated IS ISMP (one of the singony enumeration values: {*Triclinic, Monoclinic, Orthorhombic, Tetragonal, Trigonal, Hexagonal, Cubic*}).

2.3 Metabase structure

Quite reasonably, when designing integrated IS ISMP, it's required to provide search facilities for relevant information contained in other IS ISMP of distributed system. Therefore, it's required to develop some active data store that should “know” what information is contained in every integrated IS ISMP. Considering chemical objects hierarchy, some database should exist that describes information contained in integrated resources in terms of chemical systems, substances and crystal modifications. Here we come to the metabase concept – a special database that contains metadata that describe integrated IS ISMP contents in terms of chemical objects hierarchy as well as some additional information on users and their permissions together with information required to integrate distributed IS ISMP (Fig. 2).

The metabase defines integrated IS capabilities. Its structure should be flexible enough to represent metadata on integrated ISs ISMP contents and at the same time the metabase structure should be simple and versatile to describe arbitrary data source on inorganic substances properties without exhaustive additional payload currently offered by numerous materials ontologies. Taking into consideration the fact that chemical objects and their corresponding properties description is given at different detail level in different ISs ISMP, it's important to develop metabase structure that would be suitable for description of information residing in different ISs ISMP. For example, some integrated DBs contain information on particular crystal modifications properties while others contain properties description at chemical system level. Thus, integrated ISs ISMP deal with different chemical objects situated at different chemical objects hierarchy levels. For simplicity in current paper we consider only a part of metabase structure that is devoted to chemical systems and their properties (Fig. 2). The amount of this metainformation should be enough to perform search for relevant information on systems and corresponding properties.

All tables (Fig. 2) can be logically separated into several groups according to their purpose:

- DBInfo – root table, that contains information on integrated database systems;

- DBExcludeCompatibility – table that stores exception list of ISs for relevant information search;
- UsersInfo, UsersAccess – tables that contain information on integrated system users and their access rights to integrated IS ISMP;
- SystemInfo, PropertiesInfo, DBContent – tables that describe contents of integrated IS ISMP;
- CompatibilityClasses, Compatibility, Systems2ConsiderInCompatibility – tables that contain information on accessible relevance classes and their contents (currently 3 relevance classes are used [4]).
- Meta_Systems, Meta_DBSystems, Meta_SystemsHierarchy, Meta_SystemsElement – tables to describe all chemical systems contained within integrated IS ISMP with respect to their relation to each other and chemical elements, they consist of.
- Versions – service table (not shown on diagram). It is used for database schema update and versioning.

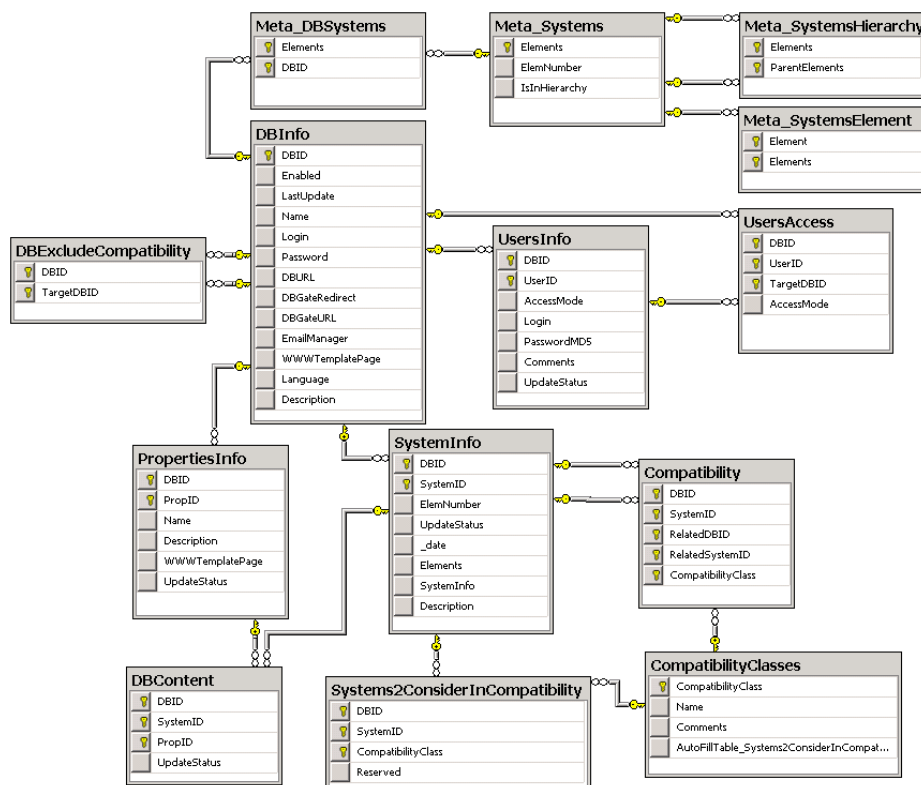


Fig. 2. A Part of the metabase logical structure on the chemical systems level only.

Taking into account chemical objects hierarchy description, a special method was developed to search for relevant information in the context of an integrated information system, based on a set-theoretic approach [5].

3 A set-theoretic approach to relevance evaluation

Relevance itself and its notion to information search is a philosophic term, covered in numerous publications. A comprehensive review of relevance itself is given by Tefko Saracevic [6]. We consider information search relevance in application to integrated IS ISMP, that area is close to “chemical similarity” [9]. So, considering chemical objects hierarchy description, a special method was developed to search for relevant information in the context of an integrated IS ISMP, based on a set-theoretic approach [5]. The main essence of set-theoretic approach is in the use of abovementioned metabase structure, that is a special database that contains information on integrable IS ISMP (set D), chemical systems (set S) and their properties (set P). To describe the relationship between the elements of the sets D , S , and P , the ternary relation W was defined on the set U (universum), which is the Cartesian product: $U = D \times S \times P$. The element (d, s, p) belongs to the relation W , where $d \in D$, $s \in S$, $p \in P$ is interpreted as follows: “the integrable IS ISMP d contains information on the p property of the chemical system s ”.

Thus, according to accepted notation the search for relevant information on a particular chemical system s can be reduced to proper definition of an R relation, which is a subset of the $S \times S$ Cartesian product (in other words, $R \subset S^2$). Thus, for any pair $(s_1, s_2) \in R$, we can state that the s_2 system is relevant to the s_1 system. For the practical solution of the problems of searching for relevant information in integrable information systems, the following rules are often used to construct R [3]:

1. For any set $s_1 \in S$, $s_2 \in S$, which includes the notation of chemical elements e_{ij} , $s_1 = \{e_{11}, e_{12}, \dots, e_{1n}\}$, $s_2 = \{e_{21}, e_{22}, \dots, e_{2n}\}$, it's true, that if $s_1 \subseteq s_2$ (that is, all chemical elements from s_1 system are contained in s_2 system), then $(s_1, s_2) \in R$.
2. The relation R is symmetric. In other words, for any $s_1 \in S$, $s_2 \in S$ it is true that, if $(s_1, s_2) \in R$, then $(s_2, s_1) \in R$.

It should be noted that abovementioned automatic variant of R relation generation is just one of the simplest and most obvious variants of such rules, and in fact more complex mechanisms can be used to get R relation. Other alternatives are used to build the R relation, called *relevance classes*. For example, browsing information on a particular property of a compound in one of integrated IS ISMP (in fact, it is information defined by (d_1, s_1, p_1) triplet), we consider (d_2, s_2, p_2) triplet to be relevant information. (d_2, s_2, p_2) triplet characterizes information on some other property of a chemical system from another integrated IS ISMP. This enables us to define relevant information more precisely, e.g. if we consider the R relations in the form: $R \subset (d_1, s_1, p_1) \times (d_2, s_2, p_2)$, where $d_1, d_2 \in D$, $s_1, s_2 \in S$, $p_1, p_2 \in P$. Actually, it's possible to even define a set of several R relations (R_1, R_2, \dots, R_n) by applying different rules to enable users to perform search for relevant information based on wide variety of R interpretations. However complex interpretations of R ($R \subset (d_1, s_1, p_1) \times (d_2, s_2, p_2)$) are not being currently used in IMET RAS, since metabase structure would be more complex to store such relations however its reasonability is not so clear. In IMET RAS simple relevancy relations of $R \subset S^2$ are used. More rules to form *relevance classes* are given in [4].

Improvement of the search relevance can also be achieved by using the c_i level, i.e. taking into account the quantitative composition of a substance, or crystal modifications

of a specific substance m_i instead of chemical system designations s_i in cases when a user requests relevant information, being at the level of inorganic substances or their modifications in the system-substance-modification hierarchy concepts [5].

When searching at the substance level, the quantitative compound composition is taken into account. The pair $(a_{i\min}, a_{i\max})$ denotes the quantitative inclusion of chemical element $e_i \in s$ into the composition, $a_{i\min}, a_{i\max} \in R^+$, $a_{i\min} \leq a_{i\max}$. If $a_{i\min} = a_{i\max}$, then the substance has a constant composition by the element $e_i \in s$. For each element of the chemical system $e_i \in s$, user during the search could specify a pair $(r_{i\min}, r_{i\max})$, where $r_{i\min}, r_{i\max} \in R^+$, denoting the allowable interval of the i -th element in the substance (R^+ is the set of non-negative real numbers). Then all substances belonging to the same chemical system are considered relevant, if for each pair $(r_{i\min}, r_{i\max})$ the following is correct: $a_{i\min} \in [r_{i\min}, r_{i\max}]$ or $a_{i\max} \in [r_{i\min}, r_{i\max}]$. In other words, if the logical disjunction $[r_{i\min} \leq a_{i\min} \ \& \ a_{i\min} \leq r_{i\max}] + [r_{i\min} \leq a_{i\max} \ \& \ a_{i\max} \leq r_{i\max}] = true$ for all $e_i \in s$, then the data on the substance are considered relevant.

When searching for relevant information taking into account the crystal modifications of m_i , crystal systems are taken into account, since often information on crystal structures is shown in different ways. For example, for lithium niobate (LiNbO_3) a hexagonal or trigonal crystallographic system is indicated in different information sources of the IS ISMP, which, in fact, corresponds to the same crystal modification.

However, it should be noted that despite the fact that the described approach, in general, provides an acceptable level of search relevance for inorganic compounds, it suffers from the inability to obtain a quantitative assessment of the search relevance and, as a consequence, the fundamental inability of search results changes by adjusting some parameters or corresponding metrics. Note that such an adjustment is useful in some cases, in particular when preparing training data sets for machine learning tasks in computer-aided construction of inorganic compounds [7].

4 Graph approach to relevance assessment

To search for relevant information and obtain a quantitative measure of relevance assessment within an integrated information system based on the properties of inorganic substances and materials, we propose to use a graph model based on the weighted graph $G = (V, E)$, built on chemical objects described as part of an integrated information system.

Let's define a set of vertices V for graph G . In accordance with the accepted three-level description of chemical objects in an integrated information system, the set of vertices consists of three disjoint subsets $V = \{S, C, M\}$, where S is the set of chemical systems s_i (qualitative compound composition), C is the set of chemical compounds c_i (the quantitative compound composition or the substance formula), M is the set of crystal modifications m_i of specific substances.

Define a set of edges E for graph G , as the union of non-intersecting subsets $E = E_s \cup E_c \cup E_m \cup E_{sc} \cup E_{cm}$, where E_s – edges that are incidental only to the set of vertices S ; E_c – edges that are incidental only to the set of substances C ; E_m – the edges that are incidental only to modifications set M . The vertices connectivity for the

classes of S , C , M is achieved by two sets of edges: Esc edges to connect vertices from S and C sets; and Ecm edges to connect vertices from C and M . Please note, that the edges connecting vertices from S and M sets, are absent.

To define the elements of the E subsets we need to introduce a couple of trivial functions: $Fs(c)$ and $Fc(m)$. The $Fs(c)$ function returns the chemical system for a given compound c , i.e. it allows to get qualitative composition from quantitative composition. The $Fc(m)$ function returns quantitative composition of a particular crystal modification of the substance, i.e. it allows to get quantitative composition from a particular crystal structure of the compound. Then, given that the chemical system is a set of chemical elements $s = \{e_1, e_2, \dots, e_n\}$ we get the following set of edges:

$$Es = \{(s_i, s_j)\}, \text{ where } s_i = \{e_{i1}, e_{i2}, \dots, e_{in}\}, s_j = \{e_{j1}, e_{j2}, \dots, e_{jm}\}, |s_i| = n, |s_j| = m, m - n = 1, s_i \subset s_j; \quad (1)$$

$$Ec = \{(c_i, c_j)\}, \text{ where } Fs(c_i) = Fs(c_j); \quad (2)$$

$$Em = \{(m_i, m_j)\}, \text{ where } Fc(m_i) = Fc(m_j); \quad (3)$$

$$Esc = \{(s_i, c_j)\}, \text{ where } Fs(c_j) = s_i; \quad (4)$$

$$Ecm = \{(c_i, m_j)\}, \text{ where } Fc(m_j) = c_i. \quad (5)$$

When searching for relevant information for a chemical object, it is necessary that a path should exist in graph between the corresponding object and a relevant one, and it is easy to calculate the measure of relevance by adding the weights of the edges on the corresponding path. Thus, we come to the necessity of introducing a real-valued function W defined on the set of graph edges:

$$W(Es) = 1000; \quad (2.1)$$

$$W(Ec) = W((c_i, c_j)) = \min(\sum_{k=0}^n 10^k * |q_{ik} - q_{jk}|); \quad (2.2)$$

where $n = |Fs(c_i)| = |Fs(c_j)|$, q_{ik} and q_{jk} – quantitative occurrence of k -th element at c_i and c_j compositions, i.e. $Q: e_k \rightarrow \mathbf{R}^+$ (respectively $Q(e_{ik}) = q_{ik}$, $Q(e_{jk}) = q_{jk}$), and the order of elements in substances is selected so to ensure the minimum value of the $W(Ec)$ objective function.

$$W(Em) = 0.1; \quad (2.3)$$

$$W(Esc) = 100; \quad (2.4)$$

$$W(Ecm) = 1. \quad (2.5)$$

As an example, we give a fragment of the relevance graph for chemical systems Cu-In-S and In-S (Fig. 3). On this example we emphasize its properties and justify the role of edge weights for quantitative assessment of the chemical objects' relevance.

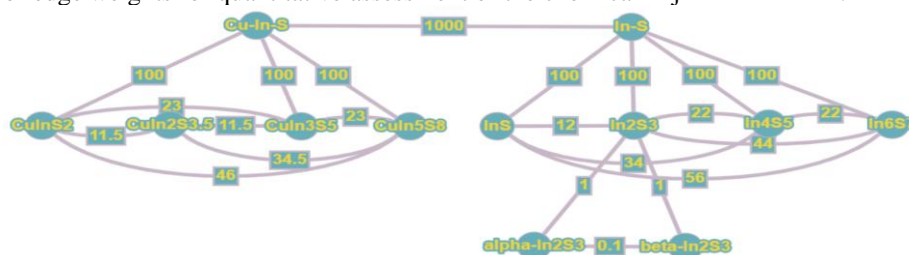


Fig. 3. Fragment of the relevance graph for Cu-In-S and In-S chemical systems.

Based on the definition of the set of edges E , it can be seen that the relevance graph is partitioned into subgraphs based on the vertices of a set S (chemical systems). Moreover, there is no path in the graph between substances from different chemical systems, bypassing the vertices of chemical systems. The vertices of the systems themselves are connected by an edge only if the set of elements of one of the systems is an own subset of the other system and their powers (i.e. a number of chemical elements that built up a system) differ by one.

Consider a subgraph constructed on the basis of the In-S chemical system vertex and consisting of substances and their corresponding modifications related to this system. It should be noted that the subgraph composed of vertices of a C set (compounds, i.e. qualitative formula) is complete, as all the vertices (InS, In₂S₃, In₄S₅, In₆S₇) are connected to each other and form a clique. Note, however, that the weights of the ribs connecting the vertex substances are different. Edge weight is a quantity characterizing the degree of closeness of corresponding quantitative compositions: the smaller the difference, the lower the weight («cost») of transition along the edge, and the corresponding substance is considered more relevant than other with greater weight of transition.

Similarly, modifications subgraph constructed on the basis of the vertex designating a particular compound is complete, and the weights of all edges are equal to 0.1. In Fig. 3 such edges are connected to each other, e.g. α -In₂S₃ and β -In₂S₃ vertices. Note, that the transition from modification to the corresponding substance has a cost of 1, and the transition from substance to the system – 100, which makes more relevant data on other modifications (including crystal structure) than the transition to the level of substances to choose another qualitative composition.

5 Discussion and further model development

The proposed graph model is an attempt to reflect the similarity degree of various chemical objects even at different representation level (system, compound, modification). In this sense, the path cost is a measure of the difference between the corresponding chemical objects, which are the vertices of the graph. The more similar the objects, the «closer» they are, meaning the path cost in the graph is less. It is worth noting that, in a broad sense, according to the definitions given in the paper, the overall relevance graph is disconnected due to the absence of a path between the vertices of chemical systems, that have no common chemical elements (i.e. $s_1 \in S$, $s_2 \in S$ such that $s_1 \cap s_2 = \emptyset$). For example, in the current model, there is no connectivity between In-S and Ga-As chemical systems, although In and Ga are similar in many ways, as far as In and Ga are elements from the same subgroup of the periodic system. In this sense, it is advisable to introduce rules for the formation of edges between similar substances and systems (in which an element from the same periodic system subgroup changes), although such an edge should have an appropriate (sufficiently large) weight comparing with analogues with common chemical elements.

As possible ways of further graph model development, one can offer the transformation of edges from the sets Esc and Ecm in pairs of arcs. In this case, the weight of the arc in the direction from the modification to the substance and from the substance

to the system should be made much less than the weight of the original edge, and the reverse arc should preserve the original edge weight. This measure will allow to obtain relevant information, described one or two levels above, which is a common way of information search in chemistry.

6 Conclusion

In the paper by means of the graph model, the concept of relevant information search was extended regarding to integrated IS ISMP. The new model allows to obtain quantitative relevance assessment of information retrieval based on the path calculation in a weighted graph, which allows ranking of chemical information found in consolidated data sources. The proposed approach is applicable not only to improve information retrieval for end users – material chemists, but also for application to computer aided design of inorganic compounds at the stage of training samples formation based on the quantitative relevance assessment.

This work was partially supported by the Russian Foundation for Basic Research (project no. 18-07-00080) and the State task № 075-00746-19-00.

References

1. Kiselyova, N.N., Dudarev, V.A., Korzhuyev, M.A.: Database on the bandgap of inorganic substances and materials, *Inorganic Materials: Applied Research*. 2016. v.7. № 1. p. 34-39.
2. Kiseleva, N.N., Prokoshev, I.V., Dudarev, V.A., Khorbenko, V.V., Belokurova, I.N., Podbelsky, V.V., Zemskov, V.S.: Database system on materials for electronics on the Internet. *Inorganic materials*, 2004, t.40, №3, p. 380-384.
3. Kornyshko, V.F., Dudarev, V.A.: Software Development for Distributed Electronics Materials. In proceedings of the Third International Conference “Information Research, Applications and Education - i.Tech”, Sofia, FOI-Commerce, 2005, pp. 27-33.
4. Dudarev, V.A., Kiselyova, N.N., Xu, Y., Yamazaki, M.: Virtual integration of the Russian and Japanese databases on properties of inorganic substances and materials. MITS 2009. In Proceedings of Symposium on Materials Database, National Institute for Materials Science (NIMS), Materials Database Station (MDBS), 2009, p. 37-48.
5. Dudarev, V.A.: Integration of information systems in the field of inorganic chemistry and materials science. ISBN 978-5-396-00745-1, M.: KRASAND, 2016, 320 p.
6. Saracevic, T.: Relevance: A review of the literature and a framework for thinking on the notion in information science. Part II: nature and manifestations of relevance. *Journal of the American Society for Information Science and Technology*, 2007, v.58(3), p.1915-1933.
7. Sen'ko, O.V., Kiselyova, N.N., Dudarev, V.A., Dokukin, A.A., Ryazanov, V.V.: Various Machine Learning Methods Efficiency Comparison in Application to Inorganic Compounds Design. In Selected Papers of the Data Analytics and Management in Data Intensive Domains. Proceedings of the XX International Conference – DAMDID / RCDL'2018, October 9-12, 2018, Moscow, V. 2277, p.152-158.
8. Serain, D. *Middleware and Enterprise Application Integration*. London: Springer-Verlag, 2002. ISBN 978-1-85233-570-0. 288 p.
9. Johnson, A.M.; Maggiora, G.M. *Concepts and Applications of Molecular Similarity*. New York: John Willey & Sons, 1990. ISBN 978-0-471-62175-1. 393 p.