УДК 004.942
DOI: 10.25559/SITITO.16.202001.13-22

О мерах и метриках релевантности информационного поиска в системах по свойствам неорганических веществ

В. А. Дударев[1,2*], И. О. Темкин[2]

[1] Национальный исследовательский университет «Высшая школа экономики», г. Москва, Россия 101000, Россия, г. Москва, ул. Мясницкая, д. 20

[2] Национальный исследовательский технологический университет «МИСиС», г. Москва, Россия 119049, Россия, г. Москва, Ленинский пр., д. 4

[*] vdudarev@hse.ru

Аннотация

Информационно-справочные системы играют серьезную роль в современном образовании, обеспечивая информационную базу для множества дисциплин. Одной из основных задач при интеграции информационных систем в учебный процесс является обеспечение релевантного поиска информации, консолидированной из гетерогенных источников. В области неорганической химии и материаловедения известны теоретико-множественные методы поиска релевантной информации, обеспечивающие построение достаточно качественного ответа на запросы пользователей. Однако проблема количественной оценки релевантности информационного поиска в этой предметной области пока остается открытой. В настоящей работе предлагается метод количественной оценки релевантности информационного поиска в интегрированных системах по свойствам неорганических веществ и материалов на основе взвешенных графов. Вершинами графа являются разносортные химические объекты (системы, вещества и кристаллические модификации) на которых определяется метрика, оценивающая степень похожести химических объектов. В метрическом пространстве определение стоимости пути между вершинами графа позволяет оценивать близость (релевантность) химических объектов, что является важным для обеспечения поиска родственных химических сущностей и их свойств в контексте интегрированной информационной системы, консолидирующей российские и зарубежные ресурсы по свойствам неорганических веществ (www.imet-db.ru). Таким образом, метрика релевантности (вводится как величина обратно пропорциональная стоимости пути в графе) позволяет оптимальным образом с точки зрения материаловеда ранжировать сведения, выводимые по запросу пользователя в единой точке доступа к консолидированным информационным ресурсам по свойствам неорганических веществ. Дополнительно к метрике на графе определяется мера, полезная при поиске полного информационного описания химического объекта. Мера используется при консолидации всех свойств объекта, доступных в интегрируемых ресурсах, что необходимо при составлении полного аналитического описания.

Ключевые слова: релевантность в неорганической химии, мера релевантности, метрика релевантности, химическое сходство, графовая модель.

Финансирование: исследование выполнено при частичной финансовой поддержке Российского фонда фундаментальных исследований в рамках научного проекта № 18-07-00080_а «Разработка алгоритмов и программ, основанных на методах распознавания образов по прецедентам, для прогнозирования количественных физических и химических свойств при конструировании еще не полученных неорганических веществ и материалов».

Для цитирования: Дударев, В. А. О мерах и метриках релевантности информационного поиска в системах по свойствам неорганических веществ / В. А. Дударев, И. О. Темкин. – DOI 10.25559/SITITO.16.202001.13-22 // Современные информационные технологии и ИТ-образование. – 2020. – Т. 16, № 1. – С. 13-22.

On Measures and Metrics of Information Retrieval Relevance in Systems on Inorganic Substances Properties

V. A. Dudarev[a,b*], **I. O. Temkin**[b]
[a] National Research University Higher School of Economics, Moscow, Russia
20 Myasnitskaya Str., Moscow 101000, Russia
[b] National University of Science and Technology "MISIS", Moscow, Russia
4 Lenin Ave., Moscow 119049, Russia
[*] vdudarev@hse.ru

Abstract

Information systems play a serious role in modern education, providing an information basis for many disciplines. One of the main tasks in integrating information systems into the educational process is to provide a relevant search for information consolidated from heterogeneous sources. In the field of inorganic chemistry and material science, set-theoretic methods for searching for relevant information are known, which ensure the construction of a sufficiently high-quality response to user requests. However, the problem of quantifying the relevance of information retrieval in this subject area remains open. In this paper we propose a method based on weighted graphs for quantifying the relevance of information retrieval in integrated systems on inorganic substances and materials properties. The vertices of the graph are heterogeneous chemical objects (systems, substances and crystal modifications) on which a metric is determined that estimates the similarity of chemical objects. In metric space, cost definition of the path between the vertices of the graph allows us to evaluate the chemical objects similarity (relevance), that is important in enabling the search for related chemical entities and their properties in the context of an integrated information system that consolidates Russian and foreign resources on inorganic substances properties (www.imet-db.ru). Thus, a relevance metric (introduced as a value inversely proportional to the cost of the graph path) allows us, from the material scientist's point of view, to optimally rank the information that is displayed at the user's request at a single access point to consolidated information resources on inorganic substances properties. In addition to the metric on the graph, a measure is defined that is useful in finding out a complete informational description of a chemical object. The measure is used in the search for all properties of the object available in integrated resources, which is necessary when compiling a complete analytical description of a chemical object.

Keywords: Relevancy in inorganic chemistry, relevance measure, relevance metric, chemical similarity, graph model.

Funding: The study was carried out with partial financial support from the Russian Foundation for Basic Research within the framework of the scientific project No. 18-07-00080_a "Development of Algorithms and Programs Founded on Precedent-Based Pattern Recognition Methods to Predict Quantitative Physical and Chemical Properties in the Design of the Still-Missing Inorganic Substances and Materials".

For citation: Dudarev V.A., Temkin I.O. On Measures and Metrics of Information Retrieval Relevance in Systems on Inorganic Substances Properties. *Sovremennye informacionnye tehnologii i IT-obrazovanie* = Modern Information Technologies and IT-Education. 2020; 16(1):13-22. DOI: https://doi.org/10.25559/SITITO.16.202001.13-22

Введение

Современные информационные технологии сделали возможным накопление и систематизацию огромного массива данных, собранных в неорганической химии. Особенно богаты информацией библиографические системы, созданные крупными издательскими компаниями[1] .Однако они не обеспечивают быстрого предметно-ориентированного поиска требуемых данных, поэтому создаются разрозненные фактографические информационные системы (ИС)[2] [1], что порождает проблему консолидации информации. Разработка и использование интегрированных ИС по свойствам веществ и материалов, консолидирующих сведения из разнородных информационных источников, является актуальным для обеспечения быстрого поиска требуемой информации со стороны участников образовательного процесса [2]. При разработке подобных систем основополагающим является способ представления данных об описываемых химических объектах и их свойствах, который, в свою очередь, определяет класс методов для обеспечения поиска релевантной информации и их функциональные возможности. Цель настоящей работы – представить новый подход для количественной оценки релевантности информационного поиска для интегрированных ИС по свойствам неорганических веществ и материалов (СНВМ), построенных на основе информационных структур, описывающих качественный и/или количественный состав вещества.

Текущее состояние проблемы

Развитие информационных технологий, появление мощных программно-аппаратных средств хранения и обработки информации стимулировали работы по созданию ИС в области неорганического материаловедения и сделали возможным их включение в образовательный процесс ведущих химико-технологических ВУЗов. В результате появилось множество узкоспециализированных ИС, созданных с учетом направлений исследований конкретных организаций-разработчиков, ориентированных на решение специфических задач. В качестве примера можно привести систему баз данных (БД), разработанных и эксплуатируемых в Институте металлургии и материаловедения им. А.А. Байкова Российской академии наук (ИМЕТ РАН): БД по фазовым диаграммам систем с полупроводниковыми фазами «Диаграмма», БД по свойствам акустооптических, электрооптических и нелинейно-оптических веществ «Кристалл», БД по свойствам тройных и четверных соединений «Фазы», БД по ширине запрещенной зоны неорганических веществ «Bandgap» и т.д., в которых хранятся данные о самых разных свойствах веществ. Эти БД отличаются не только структурами данных, но и программно-аппаратными средствами, обеспечивающими их функционирование [3-4]. Следует отметить, что они содержат обширную информацию, но в достаточно узкой области. Каталог подобных БД может быть получен по ссылке http://iric.imet-db.ru/ [5]. Ситуация, когда ни одна из разработанных систем не содержит полного набора данных о свойствах того или иного объекта (вещества или материала) и обучаемому требуется использовать сразу несколько информационных ресурсов для поиска необходимых сведений, является типичной не только для неорганического материаловедения, но и для других предметных областей [6].

Очевидно, что для обеспечения качественного информационного-образовательного сервиса для материаловедов необходима интеграция информационных систем в данной предметной области. В нашей стране первые успешные шаги в этом направлении были предприняты в начале века в ИМЕТ РАН для интеграции наиболее используемых российскими пользователями информационных систем [7]. Это позволило консолидировать информационные ресурсы и значительно сократить время, затрачиваемое пользователями на поиск нужных сведений. Примененный подход консолидации на основе метода EAI показал свою эффективность и хорошую масштабируемость при подключении к интегрированной ИС СНВМ ресурсов, разработанных в разных организациях (например, справочника по термическим константам веществ – «ТКВ», разработанного ОИВТ РАН совместно с МГУ) и даже странах (ИС «AtomWork», разработанной в NIMS, Япония) [8].

Одной из сложностей при интеграции гетерогенных ИС СНВМ является разнородность описываемых в них химических объектов. Так, например, ИС «Диаграмма» содержит сведения на уровне химической системы, т.е. набора химических элементов, образующих определенную фазовую диаграмму полупроводниковой системы. Другие ИС СНВМ (например, ИС «Кристалл») описывают свойства веществ с конкретным соотношением элементов, с учетом кристаллических модификаций веществ, т.е. на этом уровне описывается количественный состав вещества и его кристаллическая решетка. Такая несовместимость описаний химических объектов в разных ИС СНВМ диктует необходимость использования многоуровневого описания химических объектов в интегрированной ИС, как минимум, на уровне химических систем, веществ и их модификаций. С учетом подобного описания был разработан метод поиска релевантной информации в контексте интегрированной ИС СНВМ, основанный на теоретико-множественном подходе [9].

Теоретико-множественный подход к оценке релевантности

Суть теоретико-множественного подхода заключается в использовании метабазы, содержащей информацию по интегрируемым ИС СНВМ (множество D), химическим системам (множество S) и их свойствам (множество P). Для описания взаимосвязи между элементами множеств D, S и P определялось тернарное отношение W на множестве U, являющимся декартовым произведением: $U = D \times S \times P$. Принадлежность элемента (d, s, p) отношению W, где $d \in D$, $s \in S$, $p \in P$ интерпретируется следующим образом: “в интегрируемой ИС d содержится информация по свойству p химической системы s”.

[1] ACS Publications [Электронный ресурс]. URL: https://pubs.acs.org/ (дата обращения: 02.12.2019); ScienceDirect [Электронный ресурс]. URL: https://www.sciencedirect.com (дата обращения: 02.12.2019); Springer Nature [Электронный ресурс]. URL: https://link.springer.com (дата обращения: 02.12.2019); Wiley Online Library [Электронный ресурс]. URL: https://onlinelibrary.wiley.com (дата обращения: 02.12.2019).

[2] NIST Data Gateway [Электронный ресурс]. URL: https://www.nist.gov/srd (дата обращения: 02.12.2019); NIMS Materials Database (MatNavi) [Электронный ресурс]. URL: http://mits.nims.go.jp/index_en.html (дата обращения: 02.12.2019); SpringerMaterials Homepage https://materials.springer.com (дата обращения: 02.12.2019).

Таким образом, поиск релевантной информации о конкретной химической системе s сводился к определению отношения R, являющегося подмножеством декартова произведения $S \times S$ (иными словами, $R \subset S^2$). Таким образом, о любой паре $(s_1, s_2) \in R$ можно сказать, что система s_2 является релевантной системе s_1. Для практического решения задач поиска релевантной информации в интегрируемых ИС СНВМ часто использовались следующие правила для построения отношения R [7]:

1) Для любых множеств $s_1 \in S$, $s_2 \in S$, в которые входят обозначения химических элементов e_{ij}, $s_1 = \{e_{11}, e_{12}, ..., e_{1n}\}$, $s_2 = \{e_{21}, e_{22}, ..., e_{2n}\}$, верно, что, если $s_1 \subseteq s_2$ (то есть, все химические элементы из системы s_1 содержатся в системе s_2), то $(s_1, s_2) \in R$.

2) Отношение R симметрично. Иными словами, для любых $s_1 \in S$, $s_2 \in S$ верно, что, если $(s_1, s_2) \in R$, то и $(s_2, s_1) \in R$.

Также использовались альтернативные варианты построения отношения R, названные классами релевантности и отмечалась возможность более четкого определения релевантной информации при использовании отношений R вида: $R \subset (d_1, s_1, p_1) \times (d_2, s_2, p_2)$, где $d_1, d_2 \in D$, $s_1, s_2 \in S$, $p_1, p_2 \in P$.

Улучшение релевантности поиска достигалось также за счет использования уровня, учитывающего количественный состав вещества c_i или кристаллическую модификацию конкретного вещества m_i, вместо обозначений химических систем s_i в случаях, когда пользователь запрашивал релевантную информацию, находясь на уровне неорганических веществ или их модификаций в иерархии понятий система-вещество-модификация [9].

При поиске на уровне веществ учитывался количественный состав соединения. Парой $(a_{i\min}, a_{i\max})$ обозначалось количественное вхождение химического элемента $e_i \in s$ в состав, $a_{i\min}, a_{i\max} \in R^+$, $a_{i\min} \le a_{i\max}$. Если $a_{i\min} = a_{i\max}$, то вещество имеет постоянный состав по элементу $e_i \in s$. Для каждого элемента химической системы $e_i \in s$ пользователь при поиске мог задать пару $(r_{i\min}, r_{i\max})$, где $r_{i\min}, r_{i\max} \in R^+$, обозначающую допустимый интервал вхождения i-го элемента в состав вещества (R^+ – множество неотрицательных действительных чисел). Тогда релевантными являлись все вещества, относящиеся к той же химической системе, у которых для каждой пары $(r_{i\min}, r_{i\max})$ выполняется $a_{i\min} \in [r_{i\min}, r_{i\max}]$ или $a_{i\max} \in [r_{i\min}, r_{i\max}]$. Другими словами, если логическая дизъюнкция $[r_{i\min} \le a_{i\min} \,\&\, a_{i\min} \le r_{i\max}] + [r_{i\min} \le a_{i\max} \,\&\, a_{i\max} \le r_{i\max}] = true$ для всех $e_i \in s$, то данные о веществе считаются релевантными.

При поиске релевантной информации с учетом кристаллических модификаций m_i учитывались сингонии, т.к. часто информация о кристаллических структурах указывалась по-разному. Например, для ниобата лития ($LiNbO_3$) в разных информационных источниках ИС СНВМ указывается гексагональная или тригональная кристаллографическая система, что, по сути, соответствует одной кристаллической модификации.

Однако, следует заметить, что несмотря на то, что описанный подход, в целом, обеспечивает приемлемый уровень релевантности поиска, он страдает от отсутствия возможности получения количественной оценки релевантности поиска и, как следствие, принципиальной невозможности настройки параметров поисковой выдачи на основе изменения соответствующих метрик. Отметим, что такая настройка является полезной в некоторых случаях, в частности при формировании обучающих выборок для задач машинного обучения [10-11] при компьютерном конструировании неорганических веществ [12].

Графовый подход к оценке релевантности

Для поиска релевантной информации и получения количественной меры оценки релевантности в рамках интегрированной ИС по свойствам неорганических веществ и материалов предлагается использовать графовую модель, в основе которой лежит взвешенный граф $G=(V,E)$, построенный на химических объектах, описываемых в рамках интегрированной ИС СНВМ.

Определим множество вершин V графа G. В соответствии с принятым трехуровневым описанием химических объектов в интегрированной ИС множество вершин состоит из трех непересекающихся подмножеств $V=\{S, C, M\}$, где S – множество химических систем s_i (качественный состав вещества), C – множество химических систем c_i (количественный состав вещества), таких, что и M – множество кристаллических модификаций конкретных веществ m_i.

Определим множество ребер E графа G, как объединение непересекающихся подмножеств $E=Es \cup Ec \cup Em \cup Esc \cup Ecm$, где Es – ребра, инцидентные только множеству вершин S, Ec – ребра, инцидентные только множеству веществ C, Em – ребра, инцидентные только множеству модификаций M. Связность вершин графа из классов S, C, M достигается за счет ребер Esc, связывающих вершины S и C, и Ecm, связывающих вершины С и M. Заметим, что ребра, связывающие вершины S и M, отсутствуют.

Определим элементы подмножеств E, воспользовавшись тривиальными функциями $Fs(c)$ и $Fc(m)$, возвращающими химическую систему по заданному веществу (качественный состав по количественному) и количественный состав вещества по кристаллической модификации вещества соответственно. Тогда, учитывая, что система является множеством элементов $s = \{e_1, e_2, ..., e_n\}$ получим множество ребер:

$Es = \{(s_i, s_j)\}$, где $s_i = \{e_{i1}, e_{i2}, ..., e_{in}\}$, $s_j = \{e_{j1}, e_{j2}, ..., e_{jm}\}$, $|s_i| = n$, $|s_j| = m$, $m - n = 1$, $s_i \sqsupset s_j$; (1)

$Ec = \{(c_i, c_j)\}$, где $Fs(c_i) = Fs(c_j)$; (2)

$Em = \{(m_i, m_j)\}$, где $Fc(m_i) = Fc(m_j)$; (3)

$Esc = \{(s_i, c_j)\}$, где $Fs(c_j) = s_i$; (4)

$Ecm = \{(c_i, m_j)\}$, где $Fc(m_j) = c_i$. (5)

При поиске релевантной информации для химического объекта необходимо, чтобы существовал путь между соответствующим объектом и релевантным ему, а меру релевантности несложно вычислить, учитывая веса ребер на соответствующем пути. Таким образом, приходим к необходимости введения действительной функции W, определенной на множестве ребер графа:

$W(Es) = W((s_i, s_j)) = 10^N$, где $N = \max(|s_i|, |s_j|)$; (2.1)

$W(Ec) = W((c_i, c_j)) =)$; (2.2)

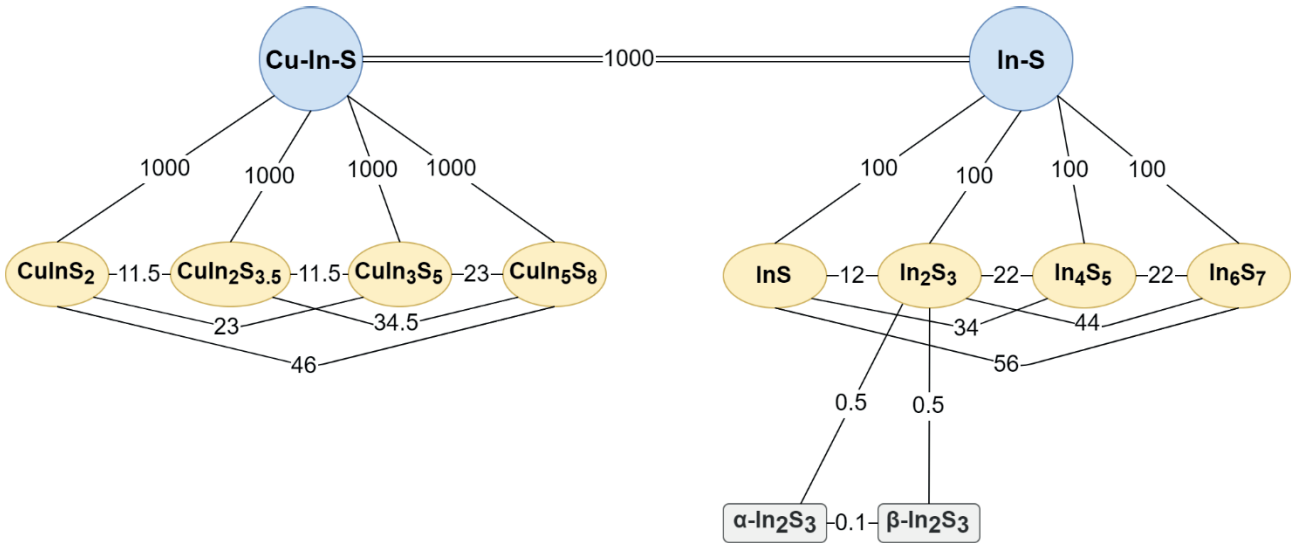
где $n = |Fs(c_i)| = |Fs(c_j)|$, q_{ik} и q_{jk} – количественное вхождение k-го элемента в состав веществ c_i и c_j, т.е. Q: $el_k \to R^+$ (соответственно $Q(el_{ik}) = q_{ik}$, $Q(el_{jk}) = q_{jk}$), а порядок элементов в веществе выбирается так, чтобы обеспечить минимальное значение целевой функции $W(Ec)$.

$W(Em) = 0{,}1$; (2.3)

$W(Esc) = W((s_i, c_j)) = 10^N$, где $N = |Fs(c_j)| = |s_i|$; (2.4)

$W(Ecm) = 0{,}5$. (2.5)

Приведем фрагмент графа релевантности для химических систем Cu-In-S и In-S (рис. 1) и на его примере укажем его свойства и обоснуем роли весов ребер для оценки релевантности химических объектов.



Р и с. 1. Фрагмент графа релевантности для систем Cu-In-S и In-S

F i g. 1. Fragment of the relevance graph for Cu-In-S and In-S systems

Исходя из определения множества ребер *E* видно, что граф релевантности разбивается на подграфы, построенные на базе вершин из множества химических систем *S*. Более того, не существует пути в графе между веществами из разных химических систем в обход вершин химических систем. Сами же вершины систем связаны ребром только в случае, если множество элементов одной из систем является собственным подмножеством другой системы и их мощности отличаются на единицу.

Рассмотрим подграф, построенный на базе вершины химической системы In-S и состоящий из веществ и их модификаций, относящихся в этой системе. Следует заметить, что подграф, образованный вершинами из множества *C*, является полным, т.к. все вершины (InS, In_2S_3, In_4S_5, In_6S_7) связаны друг с другом и образуют клику. Отметим, что отличаются веса ребер, соединяющих вещества-вершины. Вес ребра – величина характеризующая степень близости (похожести) количественных составов: чем меньше различие, тем меньше вес ("стоимость") перехода по ребру, а соответствующее вещество считается более релевантным.

Аналогично подграф модификаций, построенный на базе вершины-вещества, является полным, а веса всех ребер равны 0,1 – на рис. 1 таким ребром соединены α-In_2S_3 и β-In_2S_3. Отметим, что переход от модификации к соответствующему веществу имеет стоимость 0,5, а переход от вещества к системе – 100 (10^2), что делает более релевантным данные о других модификациях того же вещества, нежели переход на уровень других веществ.

Несложно видеть, что для функции *W* (2.1-2.5), заданной на множестве, образуемом элементами множеств *S, C и M,* выполняются аксиомы тождества, симметрии и правило треугольника. Таким образом, функция *W* является *метрикой* в пространстве химических объектов.

Особенностью ИС СНВМ является то, что данные по тем или иным свойствам веществ часто хранятся на разных уровнях детализации описания химических объектов. Например, в ИС «Кристалл» данные по показателям преломления веществ хранятся на уровне кристаллических модификаций (*M*), а сведения по температуре плавления задаются на уровне количественного состава веществ (*C*). В тоже время, данная ИС не описывает фазовые диаграммы соответствующих систем, но соответствующая дополнительная информация может быть получена из ИС «Диаграмма», которая описывает данное свойство на уровне качественного описания химических объектов, т.е. на уровне химических систем (*S*). Таким образом, для того, чтобы предоставить исчерпывающие свойства химического объекта необходимо просмотреть свойства, заданные на более высоких уровнях описания. Поэтому ребра, соединяющие вершины разных классов (*Esc* и *Ecm*) необходимо преобразовать в пары дуг с разным весом так, чтобы веса дуг от модификаций к веществам и от веществ к системам оказались равны нулю:

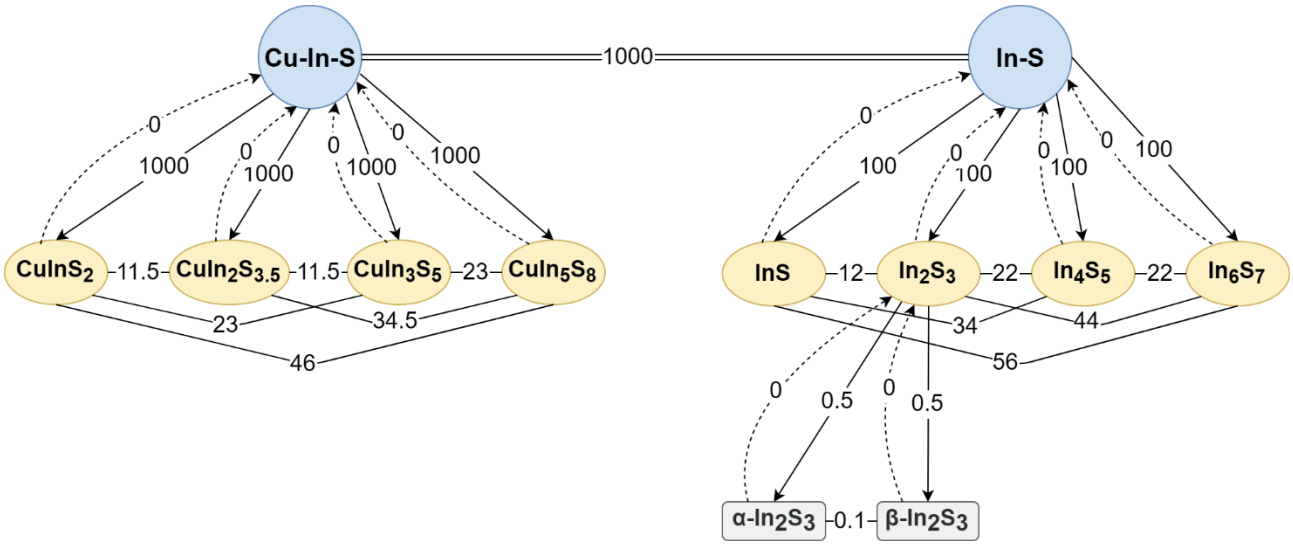
$W'(\{(c_i, m_j)\}) = 0{,}5$, где $Fc(m_j) = c_i$; (3.1)

$W'(\{(m_j, c_i)\}) = 0$, где $Fc(m_j) = c_i$; (3.2)

$W'(\{(s_i, c_j)\}) = 10^N$, где $Fs(c_j) = s_i$, $N = |\, Fs(c_j)\,|$; (3.3)

$W'(\{(c_j, s_i)\}) = 0$, где $Fs(c_j) = s_i$. (3.4)

Очевидно, что функция *W'* (3.1-3.4) является *мерой* в пространстве химических объектов, позволяющей выявлять полный набор свойств химического объекта – для этого достаточно получить свойства всех объектов, достижимых их текущего со стоимостью пути равной нулю. *W' не является метрикой* в силу нарушения симметричности и неравенства треугольника. Ориентированный граф релевантности химических свойств (рис. 2) содержит дуги с нулевым весом, обозначенные пунктиром, добавленные согласно правилам (3.2) и (3.4).



Р и с. 2. Фрагмент ориентированного графа релевантности свойств для химических систем Cu-In-S и In-S

F i g. 2. Fragment of a directed graph of the properties relevance for the chemical systems Cu-In-S and In-S

Обсуждение результатов и дальнейшего развития модели

Предложенная графовая модель является попыткой отразить степень похожести химических сущностей. Похожие попытки в областях, связанных с органической и неорганической химией, предпринимались в работах [13-20]. В этом смысле, стоимость пути, задаваемая *W* является метрикой различия между соответствующими химическими объектами, являющимися вершинами графа. Чем более похожи объекты, тем они «ближе», значит стоимость пути в графе меньше. Меру и метрику, предложенные в работе, планируется использовать при ранжировании релевантной информации, отображаемой при интеграции ИС СНВМ, в частности в единой точке доступа http://www.imet-db.ru, учитывая что релевантность обратно пропорциональна стоимости пути в графе.

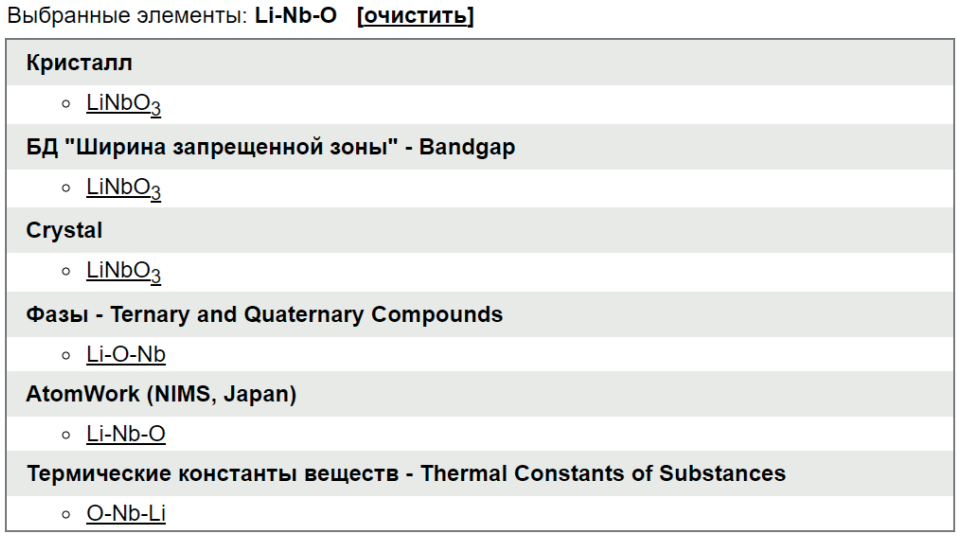
Стоит заметить, что в широком смысле, согласно данным в работе определениям, граф релевантности является несвязным из-за отсутствия пути между вершинами из , такими, что $s_1 \cap s_2 = \emptyset$. Например, в текущей модели отсутствует связность между химическими системами In-S и Ga-As, хотя In и Ga во многом похожи, как элементы из одной подгруппы периодической системы. В этом смысле в некоторых случаях с химической точки зрения целесообразно ввести правила для образования ребер между веществами и системами (в которых меняется элемент из той же подгруппы периодической системы), пусть и с достаточно большим весом.

Введение *W'* позволяет отобрать химические объекты, чтобы собрать воедино все сведения о запрашиваемой химической сущности из разных ИС СНВМ, что особенно полезно при составлении выборок для машинного обучения или при поиске всех свойств искомого химического объекта.

В качестве возможных путей дальнейшего развития графовой модели планируется уточнять веса, рассчитываемые по выражениям (2.1)-(2.5) с дальнейшим учетом многокомпонентности химических систем, т.е. принимая во внимание не только мощность множества элементов химической системы и настраивая веса дуг в подграфе системы согласно количеству химических элементов, входящих в ее состав, но и другие факторы, обусловленные наличием периодических зависимостей. Это должно способствовать более точной балансировке графа относительно рассматриваемых многокомпонентных систем.

Использование результатов работы в образовании

Прикладной аспект разрабатываемой графовой модели заключается в ее использовании в интегрированной ИС по свойствам неорганических веществ и материалов для ранжирования релевантной информации при поиске по составу соединения. Рассмотрим типовой пример использования интегрированной ИС в образовании в качестве справочника по физико-химическим свойствам веществ. Пользователь заходит на страницу www.imet-db.ru и использует таблицу Менделеева для задания качественного состава вещества, указывая, например, Li-Nb-O, при этом адрес автоматически меняется на http://www.imet-db.ru/?system=Li-Nb-O и интегрированная система выдает список найденной информации по данной химической системе (рис. 3).


Выбранные элементы: **Li-Nb-O** **[очистить]**

Кристалл
- $LiNbO_3$

БД "Ширина запрещенной зоны" - Bandgap
- $LiNbO_3$

Crystal
- $LiNbO_3$

Фазы - Ternary and Quaternary Compounds
- Li-O-Nb

AtomWork (NIMS, Japan)
- Li-Nb-O

Термические константы веществ - Thermal Constants of Substances
- O-Nb-Li

Р и с. 3. Список релевантной информации, выдаваемый интегрированной ИСЗ по системе Li-Nb-O

F i g. 3. List of relevant information issued by the integrated IC for the Li-Nb-O system

Далее пользователь выбирает интересующие его свойства, описанные в других ИС. Например, если требуются данные по ширине запрещенной зоны ниобата лития, то следует выбрать ИС «Bandgap», если диэлектрическая проницаемость, то ИС «Кристалл» (полный список свойств показывается в виде всплывающего списка при наведении на состав); если необходимы сведения по другим веществам в данной химической системе, следует выбрать ИС «Фазы» или «AtomWork». В любом случае, пользователю предоставляется возможность просмотра информации по данной химической системе и веществам, содержащейся в интегрируемых ИС, что полезно в рамках любого образовательного курса, рассматривающего свойства неорганических веществ.

Заключение

В работе с использованием графовой модели расширено понятие релевантной информации при интеграции ИС СНВМ. Новая модель позволяет получить количественные оценки релевантности информационного поиска на базе расчета пути во взвешенном графе, что позволяет осуществлять ранжирование химической информации, найденной в консолидированных источниках. При помощи весов дуг в пространстве химических объектов вводится метрика, позволяющая количественно оценивать степень похожести химических объектов. Предложенный подход применим не только для обучающихся пользователей – химиков-материаловедов, осуществляющих поиск информации, но и для использования в ИС компьютерного конструирования неорганических соединений при формировании на основе релевантности обучающих выборок для систем машинного анализа данных. Созданная интегрированная информационно-справочная система по свойствам неорганических веществ используется в учебном процессе ряда российских ВУЗов.

Список использованных источников

[1] Blokhin, E. The PAULING FILE Project and Materials Platform for Data Science: From Big Data Toward Materials Genome / E. Blokhin, P. Villars. – DOI 10.1007/978-3-319-42913-7_62-2 // Handbook of Materials Modeling / ed. by W. Andreoni, S. Yip. – Springer, Cham, 2019. – Pp. 1-25. – URL: https://link.springer.com/referenceworkentry/10.1007%2F978-3-319-42913-7_62-2#citeas (дата обращения: 02.12.2019).

[2] Абдуразаков, М. М. Что такое интеграция педагогических и информационных технологий / М. М. Абдуразаков, В. М. Монахов, М. М. Ниматулаев // Современные информационные технологии и ИТ-образование. – 2016. – Т. 12, № 4. – С. 6-10. – URL: https://www.elibrary.ru/item.asp?id=28151051 (дата обращения: 02.12.2019). – Рез. англ.

[3] Киселева, Н. Н. Компьютерные информационные ресурсы неорганической химии и материаловедения / Н. Н. Киселева, В. А. Дударев, В. С. Земсков // Успехи химии. – 2010. – Т. 79, № 2. – С. 162-188. – URL: https://www.elibrary.ru/item.asp?id=13025608 (дата обращения: 02.12.2019). – Рез. англ.

[4] Kiselyova, N. N. Integrated system of databases on the properties of inorganic substances and materials / N. N. Kiselyova, V. A. Dudarev, A. V. Stolyarenko. – DOI 10.1134/S0018151X16020085 // High Temperature. – 2016. – Vol. 54, issue 2. – Pp. 215-222. – URL: https://link.springer.com/article/10.1134/S0018151X16020085 (дата обращения: 02.12.2019).

[5] Киселева, Н. Н. База данных «Информационные ресурсы неорганической химии и материаловедения» / Н. Н. Киселева, В. А. Дударев // Информационные технологии. – 2010. – № 12. – С. 63-66. – URL: https://www.elibrary.ru/item.asp?id=15510885 (дата обращения: 02.12.2019). – Рез. англ.

[3] Поиск информации в базах данных (краткий API) // Институт металлургии и материаловедения им. А. А. Байкова Российской академии наук [Электронный ресурс]. URL: http://www.imet-db.ru/?system=Li-Nb-O (дата обращения: 02.12.2019).

[6] Pence, H. E. Big Data and Chemical Education / H. E. Pence, A. J. Williams. – DOI 10.1021/acs.jchemed.5b00524 // Journal of Chemical Education. – 2016. – Vol. 93, issue 3. – Pp. 504-508. – URL: https://pubs.acs.org/doi/10.1021/acs.jchemed.5b00524 (дата обращения: 02.12.2019).

[7] Kornyshko, V. F. Software Development for Distributed System of Russian Databases on Electronics Materials / V. F. Kornyshko, V. A. Dudarev // Information Theories & Application. – 2006. – Vol. 13, no. 2. – Pp. 121-126. – URL: http://www.foibg.com/ijita/vol13/ijita13-2-p03.pdf (дата обращения: 02.12.2019).

[8] Dudarev, V.A. Virtual integration of the Russian and Japanese databases on properties of inorganic substances and materials / V. A. Dudarev, N.N. Kiselyova, Y. Xu, M. Yamazaki // Symposium on Materials Database MITS (2009). Proceedings. Materials Database Station (MDBS). – Tsukuba, Japan, 2009. – Pp. 37-48.

[9] Дударев, В. А. Интеграция информационных систем в области неорганической химии и материаловедения. – М.: КРАСАНД, 2016.

[10] Журавлев, Ю. И. «РАСПОЗНАВАНИЕ». Математические методы. Программная система. Практические применения / Журавлев Ю. И., Рязанов В. В., Сенько О. В. – М.: изд-во ФАЗИС, 2006. 176 с.

[11] Pedregosa, F. Scikit-learn: Machine learning in Python / F. Pedregosa, G. Varoquaux, A. Gramfort [et al.] // Journal of Machine Learning Research. – 2011. – Vol. 12. – Pp. 2825-2830. – URL: https://www.jmlr.org/papers/volume12/pedregosa11a/pedregosa11a.pdf (дата обращения: 02.12.2019).

[12] Sen'ko, O. V. Various Machine Learning Methods Efficiency Comparison in Application to Inorganic Compounds Design / O. V. Sen'ko, N. N. Kiselyova, V. A. Dudarev, A. A. Dokukin, V. V. Ryazanov // CEUR Workshop Proceedings. Selected Papers of the XX International Conference on Data Analytics and Management in Data Intensive Domains (DAMDID/RCDL'2018) / ed. by L. Kalinichenko, Ya. Manolopoulos, S. Stupnikov, N. Skvortsov, V. Sukhomlin. – Moscow, 2018. – Vol. 2277. – Pp. 152-156. – URL: http://ceur-ws.org/Vol-2277/paper27.pdf (дата обращения: 02.12.2019).

[13] Park, S. H. Towards a chromatographic similarity index to establish localised quantitative structure-retention relationships for retention prediction. II. Use of Tanimoto similarity index in ion chromatography / S. H. Park, M. Talebi, R. I. J. Amos, E. Tyteca, P. R. Haddad, R. Szucs, C. A. Pohl, J. W. Dolan. – DOI 10.1016/j.chroma.2017.02.054 // Journal of Chromatography A. – 2017. – Vol. 1523. – Pp. 173-182. – URL: https://pubmed.ncbi.nlm.nih.gov/28291517/ (дата обращения: 02.12.2019).

[14] Bjørnøy, S.H. Gelling kinetics and *in situ* mineralization of alginate hydrogels: A correlative spatiotemporal characterization toolbox / S. H. Bjørnøy, S. Mandaric, D. C. Bassett, A. K. O. Åslund, S. Ucar, J.-P. Andreassen, B. L. Strand, P. Sikorski. – DOI 10.1016/j.actbio.2016.07.046 // Acta Biomaterialia. – 2016. – Vol. 44. – Pp. 243-253. – URL: https://www.sciencedirect.com/science/article/pii/S1742706116303816#! (дата обращения: 02.12.2019).

[15] Park, K. Machine learning-based chemical binding similarity using evolutionary relationships of target genes / K. Park, Y-J. Ko, P. Durai, C-H Pan. – DOI 10.1093/nar/gkz743 // Nucleic Acids Research. – 2019. – Vol. 47, issue 20. – Pp. e128. – URL: https://academic.oup.com/nar/article/47/20/e128/5557731 (дата обращения: 02.12.2019).

[16] Wassenaara, P. Chemical similarity to identify potential Substances of Very High Concern – An effective screening method / P. Wassenaara, E. Rorijea, N. Janssena, W. Peijnenburga, M. Vijver. – DOI 10.1016/j.comtox.2019.100110 // Computational Toxicology. – 2019. – Vol. 12. – Pp. 100110. – URL: https://www.sciencedirect.com/science/article/pii/S2468111319300258 (дата обращения: 02.12.2019).

[17] Vogt, M. Modeling Tanimoto Similarity Value Distributions and Predicting Search Results / M. Vogt, J. Bajorath. – DOI 10.1002/minf.201600131 // Molecular Informatics. – 2017. – Vol. 36, issue 7. – Pp. 1600131. – URL: https://onlinelibrary.wiley.com/doi/full/10.1002/minf.201600131 (дата обращения: 02.12.2019).

[18] Mutton, T. Understanding Similarities and Differences between Two Prominent Web-Based Chemical Information and Data Retrieval Tools: Comments on Searches for Research Topics, Substances, and Reactions / T. Mutton, D. D. Ridley. – DOI 10.1021/acs.jchemed.9b00268 // Journal of Chemical Education. – 2019. – Vol. 96, issue 10. – Pp. 2167-2179. – URL: https://pubs.acs.org/doi/10.1021/acs.jchemed.9b00268 (дата обращения: 02.12.2019).

[19] Skinnider, M. Comparative analysis of chemical similarity methods for modular natural products with a hypothetical structure enumeration algorithm / M. Skinnider, C. Dejong, B. Franczak, P. McNicholas, N. Magarvey. – DOI 10.1186/s13321-017-0234-y // Journal of Cheminformatics. – 2017. – Vol. 9, Article number: 46. – URL: https://jcheminf.biomedcentral.com/articles/10.1186/s13321-017-0234-y (дата обращения: 02.12.2019).

[20] Wang, Z. Improving chemical similarity ensemble approach in target prediction / Z. Wang, L. Liang, Z. Yin, J. Lin. – DOI 10.1186/s13321-016-0130-x // Journal of Cheminformatics. – 2016. – Vol. 8, Article number: 20. – URL: https://jcheminf.biomedcentral.com/articles/10.1186/s13321-016-0130-x (дата обращения: 02.12.2019).

Поступила 02.12.2019; принята к публикации 20.03.2020; опубликована онлайн 25.05.2020.

Об авторах:

Дударев Виктор Анатольевич, доцент департамента программной инженерии, факультет компьютерных наук, Национальный исследовательский университет «Высшая школа экономики» (101000, Россия, г. Москва, ул. Мясницкая, д. 20); доцент кафедры автоматизированных систем управления, Институт информационных технологий и автоматизированных систем управления, Национальный исследовательский технологический университет «МИСиС» (119049, Россия, г. Москва, Ленинский пр., д. 4), кандидат технических наук, доцент, ORCID: http://orcid.org/0000-0001-7243-9096, vdudarev@hse.ru

Темкин Игорь Олегович, заведующий кафедрой автоматизированных систем управления, Институт информационных технологий и автоматизированных систем управления, Национальный исследовательский технологический университет «МИСиС» (119049, Россия, г. Москва, Ленинский пр., д. 4), док-

тор технических наук, профессор, ORCID: http://orcid.org/0000-0001-8150-6529, igortemkin@yandex.ru

Все авторы прочитали и одобрили окончательный вариант рукописи.

References

[1] Blokhin E., Villars P. The PAULING FILE Project and Materials Platform for Data Science: From Big Data Toward Materials Genome. In: W. Andreoni, S. Yip (ed.) Handbook of Materials Modeling. Springer, Cham; 2019. p. 1-25. (In Eng.) DOI: https://doi.org/10.1007/978-3-319-42913-7_62-2

[2] Abdurazakov M.M., Monahov V.V., Nimatulaev M.M. What is the Integration of Pedagogical and Information Technologies. *Sovremennye informacionnye tehnologii i IT-obrazovanie* = Modern Information Technologies and IT-Education. 2016; 12(4):6-10. Available at: https://www.elibrary.ru/item.asp?id=28151051 (accessed 02.12.2019). (In Russ., abstract in Eng.)

[3] Kiselyova N.N., Dudarev V.A., Zemskov V.S. Computer Information Resources in Inorganic Chemistry and Materials Science. *Russian Chemical Reviews*. 2010; 79(2):145-166. (In Eng.) DOI: http://dx.doi.org/10.1070/RC2010v079n02ABEH004104

[4] Kiselyova N.N., Dudarev V.A., Stolyarenko A.V. Integrated system of databases on the properties of inorganic substances and materials. *High Temperature.* 2016; 54(2):215-222. (In Eng.) DOI: https://doi.org/10.1134/S0018151X16020085

[5] Kiselyova N.N., Dudarev V.A. "Information Resources on Inorganic Chemistry and Materials Science" Database. *Informacionnye tehnologii* = Information Technology. 2010; 12:63-66. Available at: https://www.elibrary.ru/item.asp?id=15510885 (accessed 02.12.2019). (In Russ., abstract in Eng.)

[6] Pence H.E., Williams A.J. Big Data and Chemical Education. *Journal of Chemical Education.* 2016; 93(3):504-508. (In Eng.) DOI: https://doi.org/10.1021/acs.jchemed.5b00524

[7] Kornyshko V.F., Dudarev V.A. Software Development for Distributed System of Russian Databases on Electronics Materials. *Information Theories & Application.* 2006; 13(2):121-126. Available at: http://www.foibg.com/ijita/vol13/ijita13-2-p03.pdf (accessed 02.12.2019). (In Eng.)

[8] 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. In: Symposium on Materials Database MITS (2009). Proceedings. Materials Database Station (MDBS). – Tsukuba, Japan; 2009. p. 37-48. (In Eng.)

[9] Dudarev V.A. *Integratsiya informatsionnykh sistem v oblasti neorganicheskoy khimii i materialovedeniya* [Integration of information systems in the field of inorganic chemistry and materials science]. Krasand, Moscow; 2016. (In Russ.)

[10] Zhuravlev Yu.I., Ryazanov V.V., Sen'ko O.V. «RECOGNITION». *Matematicheskiye metody. Programmnaya sistema. Prakticheskiye primeneniya* [Mathematical methods. Software system. Practical solutions]. Phasis, Moscow; 2006. (In Russ.)

[11] Pedregosa F., Varoquaux G., Gramfort A., et al. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research.* 2011; 12:2825-2830. Available at: https://www.jmlr.org/papers/volume12/pedregosa11a/pedregosa11a.pdf (accessed 02.12.2019). (In Eng.)

[12] 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: Kalinichenko L., Manolopoulos Ya., Stupnikov S., Skvortsov N., Sukhomlin V. (eds) *Selected Papers of the XX International Conference on Data Analytics and Management in Data Intensive Domains (DAMDID/RCDL'2018).* October 9-12, 2018, Moscow. *CEUR Workshop Proceedings.* 2018; 2277:152-156. Available at: http://ceur-ws.org/Vol-2277/paper27.pdf (accessed 02.12.2019). (In Eng.)

[13] Park S.H., Talebi M., Amos R.I.J., Tyteca E., Haddad P.R., Szucs R., Pohl C.A., Dolan J.W. Towards a chromatographic similarity index to establish localised quantitative structure-retention relationships for retention prediction. II. Use of Tanimoto similarity index in ion chromatography. *Journal of Chromatography A.* 2017; 1523:173-182. (In Eng.) DOI: https://doi.org/10.1016/j.chroma.2017.02.054

[14] Bjørnøy S.H., Mandaric S., Bassett D.C., Åslund A.K.O., Ucar S., Andreassen J.-P., Strand B.L., Sikorski P. Gelling kinetics and in situ mineralization of alginate hydrogels: A correlative spatiotemporal characterization toolbox. *Acta Biomaterialia.* 2016; 44:243-253. (In Eng.) DOI: https://doi.org/10.1016/j.actbio.2016.07.046

[15] Park K., Ko Y-J., Durai P., Pan C-H. Machine learning-based chemical binding similarity using evolutionary relationships of target genes. *Nucleic Acids Research.* 2019; 47(20):e128. (In Eng.) DOI: https://doi.org/10.1093/nar/gkz743

[16] Wassenaara P., Rorijea E., Janssena N., Peijnenburga W., Vijver M. Chemical similarity to identify potential Substances of Very High Concern – An effective screening method. *Computational Toxicology.* 2019; 12:100110. (In Eng.) DOI: https://doi.org/10.1016/j.comtox.2019.100110

[17] Vogt M., Bajorath J. Modeling Tanimoto Similarity Value Distributions and Predicting Search Results. *Molecular Informatics.* 2017; 36(7):1600131. (In Eng.) DOI: https://doi.org/10.1002/minf.201600131

[18] Mutton T., Ridley D.D. Understanding Similarities and Differences between Two Prominent Web-Based Chemical Information and Data Retrieval Tools: Comments on Searches for Research Topics, Substances, and Reactions. *Journal of Chemical Education*. 2019; 96(10):2167-2179. (In Eng.) DOI: https://doi.org/10.1021/acs.jchemed.9b00268

[19] Skinnider M., Dejong C., Franczak B., McNicholas P., Magarvey N. Comparative analysis of chemical similarity methods for modular natural products with a hypothetical structure enumeration algorithm. *Journal of Cheminformatics*. 2017. Vol. 9, Article number: 46. (In Eng.) DOI: https://doi.org/10.1186/s13321-017-0234-y

[20] Wang Z., Liang L., Yin Z., Lin J. Improving chemical similarity ensemble approach in target prediction. *Journal of Cheminformatics*. 2016. Vol. 8, Article number: 20. (In Eng.) DOI: https://doi.org/10.1186/s13321-016-0130-x

Submitted 02.12.2019; revised 20.03.2020; published online 25.05.2020.

About the authors:

Victor A. Dudarev, Associate Professor of the School of Software Engineering, Faculty of Computer Science, National Research University Higher School of Economics (20 Myasnitskaya Str., Moscow 101000, Russia); Associate Professor of the Department of Automated Control Systems, College of IT & Automated Control Systems, National University of Science and Technology "MISIS" (4 Lenin Ave., Moscow 119049, Russia), Ph.D. (Engineering), Associate Professor, ORCID: http://orcid.org/0000-0001-7243-9096, vdudarev@hse.ru

Igor O. Temkin, Head of the Department of Automated Control Systems, College of IT & Automated Control Systems, National University of Science and Technology "MISIS" (4 Lenin Ave., Moscow 119049, Russia), Dr.Sci. (Engineering), Professor, ORCID: http://orcid.org/0000-0001-8150-6529, igortemkin@yandex.ru

All authors have read and approved the final manuscript.