ТЕПЛОФИЗИКА ВЫСОКИХ ТЕМПЕРАТУР, 2016, том 54, № 2, с. 228–236

УДК 546:025.4.03:004.855.5

ИНТЕГРИРОВАННАЯ СИСТЕМА БАЗ ДАННЫХ ПО СВОЙСТВАМ НЕОРГАНИЧЕСКИХ ВЕЩЕСТВ И МАТЕРИАЛОВ*

Н. Н. Киселева, В. А. Дударев, А. В. Столяренко

Федеральное государственное бюджетное учреждение науки
Институт металлургии и материаловедения им. А.А. Байкова РАН, Москва
E-mail: kis@imet.ac.ru

Поступила в редакцию 26.01.2015 г.

Разработана интегрированная система баз данных по свойствам неорганических веществ и материалов, объединяющая в настоящее время базы данных ИМЕТ РАН и базу “AtomWork” по свойствам неорганических веществ, разработанную в National Institute for Materials Science (Япония). Данная система предназначена для информационного обслуживания специалистов и для компьютерного конструирования новых неорганических соединений, для чего была разработана информационно-аналитическая система. Приведены результаты применения этой системы для конструирования еще не полученных соединений.

DOI: 10.7868/S0040364416020083

ВВЕДЕНИЕ

Увеличение количества баз данных (БД), содержащих обширные и разнообразные сведения о свойствах неорганических веществ и материалов, обусловлено ростом объема информации, а также появлением и широким использованием сравнительно недорогих и компактных компьютерных устройств, сети Интернет и удобных и доступных программных систем управления базами данных (СУБД). Компьютерные БД во многом позволяют решить проблему оперативного получения необходимой информации. За последнее десятилетие количество БД в области неорганической химии и материаловедения возросло более чем в три раза. Традиционное первое место по количеству БД занимают информационные системы с данными о теплофизических свойствах (см., например, [1–10]). Широкое использование БД с кристаллографической и кристаллохимической информацией способствовало их интенсивной разработке [11–13]. В последние годы наблюдается устойчивый рост количества БД, содержащих сведения о механических свойствах (прочности, усталости, ползучести и т.д.) неорганических веществ и материалов [14, 15]. Подробная информация о БД по свойствам неорганических веществ и материалов (БД СНВМ) приведена в [16] и справочной информационной системе IRIC (Information Resources of Inorganic Chemistry) [17, 18].

БД СНВМ широко используются в фундаментальных и прикладных исследованиях и в промышленности, однако ни одна из разработанных информационных систем не может дать исчерпывающей информации обо всей совокупности свойств конкретного вещества или материала. Часто специалисты вынуждены просматривать десятки БД, чтобы найти необходимые им значения параметров заданного вещества. Для обеспечения релевантного и быстрого поиска данных о конкретном веществе из разных информационных систем предложено использовать виртуальную интеграцию БД СНВМ. Термин “виртуальная” означает, что данные собираются не в одном хранилище данных, а находятся в организациях-разработчиках. Такой путь обеспечивает наиболее эффективное решение проблемы предоставления пользователям полной совокупности данных о конкретном неорганическом веществе из разных БД, находящихся в различных организациях и странах, созданных с использованием разных программных и аппаратных средств.

Помимо основной функции БД – информационного обслуживания специалистов – эти системы предоставляют значительно более широкие возможности оперирования данными. В частности, БД с теплофизическими и кристаллоструктурными данными широко применяются для расчетов. Примерами систем, объединяющих БД с программами термодинамических расчетов, могут служить системы F*A*C*T [2], MALT2 [5], MTDATA [6], Thermo-Calc [7] и т.п. Характерной особенностью таких систем является наличие корректного аналитического описания зависимостей. Чтобы провести расчеты, специалисту достаточно только подставить необходимые сведения, извлеченные из базы данных, в одну из выбранных моделей.

* По материалам XIV Российской конференции по теплофизическим свойствам веществ (РКТС-14). Казань. Октябрь 2014 г.

Однако большинство задач, решаемых в химической, и особенно материаловедческой, практике (прогноз новых веществ с заданными свойствами, расшифровка спектральной информации, отбор веществ в качестве материалов для применения в определенных целях, разработка оптимальных технологий получения материалов, разделение и идентификация веществ и т.п.), плохо поддаются формализации на уровне тех сравнительно простых алгебраических структур, которые используются, например, термодинамикой. Сейчас все эти проблемы чаще всего решаются на основе использования эмпирических знаний. Для решения таких плохо формализуемых задач создаются информационно-аналитические системы (ИАС), объединяющие БД СНВМ и подсистемы анализа информации.

ПРИНЦИПЫ ИНТЕГРАЦИИ БД СНВМ

Принципиально возможны два подхода к интеграции баз данных. В первом, основанном на концепции общего хранилища данных, информация из разных БД после унификации и очистки от явных ошибок загружается в мегабазу данных – Data Warehouse. Этот подход является основой для методологии интеграции ETL (Extract, Transform, Load) – программного обеспечения для извлечения и преобразования информации из интегрируемых баз данных и последующей загрузки данных в хранилище [19]. Однако специфика предметной области делает полное слияние информации БД СНВМ и создание некого центрального хранилища данных крайне сложной технической и организационной задачей, требующей огромных финансовых вложений.

В связи с этим второй подход – виртуальная интеграция БД СНВМ и создание неоднородной распределенной информационной системы – является более эффективным и менее затратным, обеспечивающим независимость развития отдельных БД и позволяющим организовать доступ ко всему массиву данных о конкретном веществе или материале из разных БД.

При виртуальной интеграции БД возможны два основных технологических приема: 1) интеграция корпоративной информации (Enterprise Information Integration – EII) [20], 2) интеграция корпоративных приложений (Enterprise Application Integration – EAI) [21].

При интеграции информации (EII) разрабатывается программный интерфейс доступа, с помощью которого можно извлекать необходимые данные из разных БД. Строится некая центральная информационная система (метабаза), которая взаимодействует с распределенными источниками данных, извлекает и предоставляет пользователю агрегированную информацию о запрашиваемом веществе из разных БД СНВМ.

Использование второго подхода (EAI) наиболее целесообразно, когда БД СНВМ включают прикладные программы. При реализации этого подхода объединяются не сами БД, а только их пользовательские интерфейсы, которые осуществляют доступ к расчетным подсистемам. Такими интерфейсами могут быть web-приложения соответствующих информационных систем. Следует отметить, что при больших объемах данных и сложных расчетах в этом подходе уместно использование “облачных” вычислений.

Анализ возможностей существующих технологий интеграции БД с точки зрения информационных потребностей специалистов в области неорганической химии и материаловедения показал, что ни один из вышеуказанных подходов по отдельности не способен решить все проблемы, возникающие при интеграции информационных систем и программных приложений. В связи с этим при создании интегрированной системы БД СНВМ был предложен комплексный подход, сочетающий в себе интеграцию на уровне данных и пользовательских интерфейсов (EII + EAI) [22, 23]. С одной стороны, он позволил провести интеграцию пользовательских интерфейсов web-приложений БД СНВМ и прикладных программ анализа данных (EAI), а с другой – обеспечил консолидацию информации, извлеченной из разнородных распределенных источников данных (EII).

ИНТЕГРИРОВАННАЯ СИСТЕМА БД СНВМ ИМЕТ РАН

Предложенный подход к интеграции был успешно применен при создании интегрированной системы БД СНВМ, которая в настоящее время объединяет информационные системы, разработанные в ИМЕТ РАН [16]: по фазовым диаграммам полупроводниковых систем (“Диаграмма”), свойствам акустооптических, электрооптических и нелинейнооптических веществ (“Кристалл”), ширине запрещенной зоны неорганических веществ (“Bandgap”), свойствам неорганических соединений (“Фазы”) и свойствам химических элементов (“Elements”), а также БД “AtomWork” по свойствам неорганических веществ, разработанную в National Institute for Materials Science (NIMS, Япония) (рис. 1).

БД по свойствам неорганических соединений “Фазы” [16, 24] в настоящее время содержит информацию о свойствах около 52 тысяч тройных соединений и более 31 тысячи четверных соединений, собранную более чем из 32 тыс. литературных источников. Она включает краткую информацию о наиболее распространенных свойствах неорганических соединений: кристаллохимических (тип кристаллической структуры с указанием температуры и давления, выше которых реализуется указанная структура, сингония, пространственная группа, число формульных единиц в

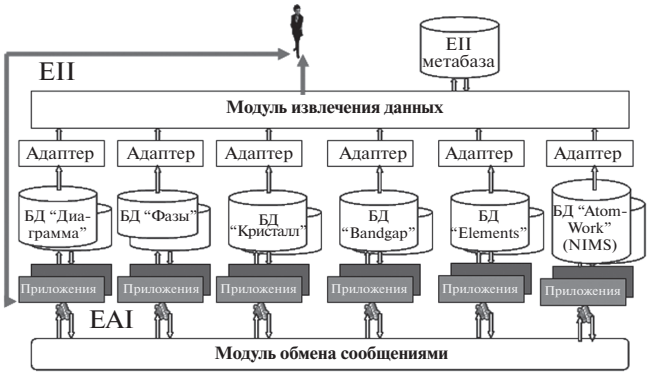

Рис. 1. Структура интегрированной системы баз данных по свойствам неорганических веществ и материалов ИМЕТ РАН.

элементарной ячейке, параметры кристаллической решетки) и теплофизических (тип и температура плавления, температура распада соединения в твердой или газообразной фазах и температура кипения при атмосферном давлении). Помимо этого, БД содержит информацию о сверхпроводящих свойствах соединений. БД "Фазы" формируется на основе анализа сведений, почерпнутых из периодических изданий, справочников, монографий, отчетов, а также реферативных журналов (более половины источников хранятся в виде pdf-документов). БД "Фазы" доступна зарегистрированным пользователям в сети Интернет [25].

БД "Elements" [16] включает информацию о 90 наиболее распространенных свойствах химических элементов: теплофизических (температура плавления и кипения при 1 атм, стандартные теплопроводность, молярная теплоемкость, энтальпия атомизации, энтропия и т.д.), размерных (ионные, ковалентные, металлические, псевдопотенциальные радиусы, объем атома и т.д.), других физических свойствах (магнитной восприимчивости, электропроводности, твердости, плотности и т.д.), положении в Периодической таблице элементов и т.д. БД доступна в сети Интернет [26].

БД "Диаграмма" [16, 27, 28] содержит собранную и оцененную высококвалифицированными экспертами информацию, о десятках фазовых *P*, *T*, *x*-диаграмм двух- и трехкомпонентных полупроводниковых систем и о физико-химических свойствах образующихся в них фаз. Рисунки с фазовыми диаграммами хранятся в виде растровых (формат jpeg) и векторных (формат swf) изображений. Следует отметить, что применение только стандартных программных средств со стороны пользователей является особенностью всех БД СНВМ, разработанных в ИМЕТ РАН. Помимо графической информации в подробно написанных аналитических обзорах и специальных таблицах БД хранятся расчетные модели, полученные в результате термодинамического согласования или статистической оптимизации фазовых диаграмм или их элементов с применением различных методик. Пользователь может работать с ними для своих расчетов, подставив параметры, включенные в аналитические обзоры или хранящиеся в соответствующих таблицах термодинамических свойств и описаний расчетных моделей БД. Помимо этого, в соответствующих таблицах БД хранятся оцененные экспертами и согласованные числовые данные для элементов фазовых диаграмм. БД доступна зарегистрированным пользователям в сети Интернет [29].

БД "Bandgap" [30] включает информацию о ширине запрещенной зоны более трех тыс. неорганических веществ и доступна в сети Интернет [31].

БД "Кристалл" [16, 28] включает информацию о свойствах: пьезоэлектрических (пьезоэлектрические коэффициенты, упругие постоянные и т.д.), нелинейно-оптических (нелинейно-оптические коэффициенты, компоненты тензора Миллера и т.д.), кристаллохимических (тип кристаллической структуры, сингония, пространственная и точечная группа, число формульных единиц в элементарной ячейке, параметры кристаллической решетки), оптических (показатели преломления, область прозрачности и т.д.), теплофизических (температура плавления, теплоемкость, теплопроводность и т.д.) – более 140 акустооптических, электрооптических и нелинейно-оптических веществ, собранную и оцененную высококвалифицированными экспертами в данной предметной области. Она имеет русско- и англоязычную версии, доступные зарегистрированным пользователям в сети Интернет [32].

БД Inorganic Material Database – AtomWork (NIMS, Япония) [33], содержащая информацию о более чем 82 тыс. кристаллических структур, 55 тысячах значений свойств материалов и 15 тысячах фазовых диаграмм, доступна в сети Интернет [34].

Для интеграции БД была использована SOA (сервисно-ориентированная архитектура), базирующаяся на применении web-сервисов для обеспечения взаимодействия между гетерогенными информационными системами. Для поиска релевантной информации в контексте информационных систем используется специально разработанная метабаза [16], описывающая содержимое интегрируемых БД в терминах формализованной иерархии понятий, присущих неорганической химии и материаловедению.

Интеграция БД является современной тенденцией развития информационных систем в материаловедческих областях. Интегрированные системы наиболее эффективно позволяют обеспечить специалистов достоверными и полными данными о свойствах веществ и материалов и до-

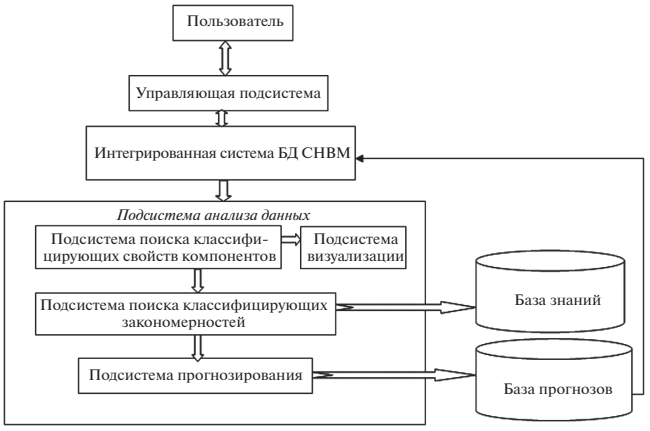

Рис. 2. Схема информационно-аналитической системы для конструирования неорганических соединений.

ставить эту совокупную информацию в любую точку мира по сети Интернет.

Дальнейшее развитие интегрированных систем БД СНВМ непосредственно связано с решением таких проблем, как:

– разработка общих стандартов и типового программного обеспечения для интеграции БД СНВМ;

– дальнейшее структурирование предметных областей;

– создание тезаурусов;

– оценка достоверности данных;

– решение организационных вопросов доступа к БД разных учреждений и разных стран.

ИАС ДЛЯ КОМПЬЮТЕРНОГО КОНСТРУИРОВАНИЯ НЕОРГАНИЧЕСКИХ СОЕДИНЕНИЙ

Одной из тенденций в разработке информационных систем является включение в их состав средств анализа данных: от простейших способов агрегации информации, обработки сложных многокритериальных запросов, программ статистического анализа и визуализации результатов до сложных систем искусственного интеллекта. Разработанная авторами специальная ИАС помимо информационного обслуживания специалистов предназначена для поиска закономерностей в больших массивах химических данных и компьютерного конструирования неорганических соединений [16, 35, 36]. Она включает (рис. 2) наряду с интегрированной системой БД СНВМ подсистему анализа информации и прогнозирования, объединяющую комплекс программ распознавания образов по прецедентам, базу найденных закономерностей (базу знаний), базу полученных прогнозов возможности образования и свойств еще не полученных неорганических соединений и управляющую подсистему.

Подсистема анализа данных

Подсистемы поиска классифицирующих закономерностей и прогнозирования. Важнейшей задачей разработки этой подсистемы был отбор математических методов, наиболее подходящих для поиска закономерностей в химических данных. Как правило, решение этой задачи выполняется методом проб и ошибок. При отборе методов распознавания образов для анализа химической информации учитывался многолетний опыт применения этих методов для конструирования неорганических соединений [37]. В результате были выбраны следующие методы и программы:

– широкий класс алгоритмов многофункциональной системы “Распознавание”, разработанной в ВЦ РАН [38] и объединяющей помимо широко известных методов также алгоритмы распознавания (основанные на вычислениях оценок), голосования по тупиковым тестам, голосования по логическим закономерностям, статистического взвешенного голосования и т.д.;

– система обучения ЭВМ процессу формирования понятий ConFor [39], в основу которой положена оригинальная организация данных в памяти ЭВМ в виде растущих пирамидальных сетей.

Выбор вышеуказанных программ анализа данных обусловлен:

1) универсальностью относительно размерностей решаемых задач: вышеуказанные системы дают возможность решения как задач прогноза редких или уникальных событий, явлений или процессов, когда начальная (обучающая) информация мала (десятки прецедентов), так и задач больших размерностей (десятки тысяч прецедентов);

2) универсальностью относительно типа данных (допускаются числовые, бинарные и номинальные признаки);

3) возможностью обработки неполной и частично противоречивой информации, когда значения некоторых признаков неизвестны или известны приближенно.

Как правило, заранее невозможно указать, какой алгоритм является наиболее эффективным при решении конкретной задачи. В связи с этим перспективным является использование методов распознавания коллективами алгоритмов. При синтезе коллективного решения во многих случаях удается компенсировать возможные ошибки распознавания отдельных алгоритмов правильными ответами других алгоритмов. Исходя из этого, в разработанную ИАС включены программы, реализующие разные стратегии принятия коллективных решений: метод Байеса; методы, использующие области компетенции, шаблоны принятия решений, логическую коррекцию; ме-

тод выпуклого стабилизатора; динамический метод Вудса; комитетные методы и т.д. [38].

Вышеуказанные алгоритмы распознавания образов и принятия коллективных решений основаны на различных принципах и используют разные формы представления искомых закономерностей (алгебраические функции, логические выражения, нейронные или растущие пирамидальные сети и т.д.). Для интеграции таких разнородных программ была использована SOA, которая позволила учесть различия в данных и информационных структурах, используемых в интегрируемых программах, а также сложные механизмы их взаимодействия. Она обеспечила возможность достаточно простого добавления новых программ анализа данных в подсистему поиска закономерностей. При интеграции приложений вместо специализированных интерфейсов между отдельными программами применена связующая среда, которая играет роль универсального программного ядра, соединяющего все приложения [40]. Преимуществом используемой технологии на основе интегрирующей среды является в первую очередь простота поддержки и расширения разработанной на ее основе системы.

Исходная информация для компьютерного анализа извлекается из интегрированной системы БД СНВМ и имеет форму таблицы в формате Excel. Каждая строка в ней содержит набор значений свойств компонентов (элементов или более простых соединений) уже известных соединений с указанием класса, к которому принадлежит это соединение (например, тип кристаллической структуры при нормальных условиях, положение температуры перехода соединения в сверхпроводящее состояние относительно 4.2 K).

При прогнозировании еще не полученных неорганических соединений и оценке их свойств используется информация только о свойствах компонентов. Процедура прогнозирования полностью автоматизирована. Пользователь указывает только обозначения компонентов, а подсистема прогнозирования сама формирует выборку для распознавания, подставляет полученные значения свойств компонентов в классифицирующую закономерность и вычисляет результат прогноза.

Подсистема поиска классифицирующих свойств компонентов. Для отбора информативных свойств компонентов химических соединений в ИАС были включены программы, основанные на алгоритмах [41–43]. Отбор свойств компонентов, наиболее информативных для классификации веществ, имеет двоякое значение. С одной стороны, удается резко сократить объем анализируемой информации, которая для многокомпонентных веществ включает сотни значений свойств элементов и более простых соединений, а также функции от этих свойств. С другой стороны, выбор свойств компонентов, наиболее важных для классификации химических веществ, дает возможность физической интерпретации полученных классифицирующих закономерностей, что повышает доверие к полученным прогнозам и позволяет найти существенные причинно-следственные связи между параметрами объектов и разработать физические и химические модели явлений.

Подсистема визуализации. Интерпретации полученных результатов способствует подсистема визуализации, которая строит проекции расположения точек, соответствующих соединениям, в двухмерных пространствах свойств компонентов, включающих не только исходные параметры, но и указанные пользователем алгебраические функции от этих параметров.

База знаний. База знаний содержит полученные классифицирующие закономерности. При ее программной реализации возникла проблема, связанная с тем, что форма представления знаний в используемых методах обучения ЭВМ существенно отличается. В связи с этим было предложено новое программное решение для хранения полученных закономерностей, а также сопутствующей информации о параметрах программ и исследуемых объектов [40]. Хранение этой информации реализовано средствами SQL-сервера и файловых структур на дисках сервера. На сервере хранятся полученные закономерности в специальном внутреннем формате программ анализа данных, а в таблицах БД на SQL-сервере – служебная информация об этих закономерностях, а именно: уникальный идентификатор закономерности, обозначение прогнозируемой характеристики, формульный состав химических соединений, обозначения свойств компонентов, используемых для описания веществ, пути к файлам на дисках, фамилия специалиста, проводившего оценку данных для обучения и поиск закономерностей, дата формирования закономерности и т.д.

База прогнозов. В базе прогнозов содержатся результаты предыдущих компьютерных экспериментов, а также ссылки на служебную информацию, хранящуюся в базе знаний. Использование базы прогнозов позволило повысить функциональность БД по свойствам неорганических веществ и материалов ИМЕТ РАН за счет предоставления пользователю не только известных сведений об уже изученных веществах, но и прогнозов еще не полученных неорганических соединений и оценок их свойств. В настоящее время идет заполнение этой базы.

Управляющая подсистема. Управляющая подсистема организует вычислительный процесс и осуществляет взаимодействие между функциональными подсистемами ИАС, а также обеспечивает доступ к системе в сети Интернет. Помимо этого, управляющая подсистема предоставляет

Прогноз возможности образования соединений состава AB_3X_3

X≡S

A / B	B	Al	P	Sc	Ti	Cr	Fe	Ga	As	Y	Rh	In	Sn	Sb	La	Ce	Pr	Nd	Pm	Sm	Eu	Gd	Tb	Dy	Ho	Er	Tm	Yb	Lu	Bi
Li	#1		1	1	#1	1	1	#2	#1			#2		#1	2	2	2	2												#2
Na	#1	#1	1	1	1	1	#1	#2	#1	1	#1	#1	1	#1					1	1	1	1	1	1	1	1	1	1	1	#1
K	#1	1	1	1	1	1	©	#2	#1	1	#1	1	1	#1	1		1	1	1	1	1	1	1	1	1	1	1	1	1	#2
Cu	1	#2	#1	#1		#2	#1		#1	#1	1	#2	#1	#1	#2	#2	2	#2			#1	#1	#1	#1	#1	#1	#1	#1	#1	#1
Rb	#1	1	1	1	1	1	©		1	1	1	#1	1	©	1			1	1	1	1	1	1	1	1	1	1		1	#1
Ag		#2					#2	#2	#1			#2	#2	#1	#2	#2	#2	#2	2	2	2		2	#2	2					#2
Cs	#1	1	1	1	1	1	#1	#2			1	1	1	#2	1			1	1	1	1	1	1	1	1	1	1	1	1	
Tl	#1		#2	1	1	#1	1	Θ	#1		1	#1	#1	#1	#2	2	#2	2	2	2	2							1	1	

X≡Se

A / B	B	Al	P	Sc	Ti	Cr	Fe	Ga	As	Y	Rh	In	Sn	Sb	La	Ce	Pr	Nd	Pm	Sm	Eu	Gd	Tb	Dy	Ho	Er	Tm	Yb	Lu	Bi
Li	1		1	1	1	1	1	2	#2	1	1	2			2	2	2	2					1	1	1	1	1	1	1	
Na	1	#1	1	1	1	1	#1	#1	#1	1	1	1	1	#2									1	1	1	1	1	1	1	1
K	1	#1	1	1	1	1	#1	#1	#1	1	1	1	©	#1		1			1	1	1	1	1	1	1	1	1	1	1	#1
Cu	1		1	#1	1	1	1	#2	#1	#1	1	#2	#1	#1	2	2	2	2		#1	#1	#1	#1	#1	#1	#1	#1	#1	#1	#1
Rb	1	1	1	1	1	1	1	1	1	1	1	1	1	©	1	1			1	1	1	1	1	1	1	1	1	1	1	#1
Ag	1		2			2	2	#2	#1	2		Θ		#2	Θ	Θ	Θ	2	2	Θ	Θ		Θ	2	2	2	2	2	2	Θ
Cs	1	1	1	1	1	1	1	#1	1	1	1	1	1	©	1				1	1	1	1	1	1	1	1	1	1	1	#1
Tl	#1		1	1	1	#1	1	Θ	#1	1	1	#2	#1	#1	2	2	2	2							1	1	1	1	1	#2

X≡Te

A / B	B	Al	P	Sc	Ti	Cr	Fe	Ga	As	Y	Rh	In	Sn	Sb	La	Ce	Pr	Nd	Pm	Sm	Eu	Gd	Tb	Dy	Ho	Er	Tm	Yb	Lu	Bi
Li	1			1	1		1	2		1	1				2	2	2	2	2						1	1	1	1	1	
Na	1	#1	1	1	1	1	1		1	1	1	1	1	#1				1	1	1	1	1	1		1	1	1	1	1	1
K	1	#1	1	1	1	1	1	#1	1	1	1	1	#1	#1	1				1	1	1	1	1	1	1	1	1	1	1	#1
Cu	1	#2		1	1	1	1	#2	#2	#1	1	#2		#2	2	2	2	2	2	#1	#1	#1	#1	#1	#1	#1	#1	#1	#1	#2
Rb	1	1	1	1	1	1	1	1	1	1	1	1	©	1	1	1			1	1	1	1	1	1	1	1	1	1	1	1
Ag	1		2	2		2	2	#2	2	2		#2	Θ	#2	2	2	2	2	2	2	2	#2	Θ	Θ	2	2	2			Θ
Cs	1	#1	1	1	1	1	1	1	1	1	1	1	©	1	1	1			1	1	1	1	1	1	1	1	1	1	1	1
Tl	1			1	1		1	Θ			1	Θ	#1	#2	#2	2	2	Θ	2											#2

пользователю программные средства для подготовки данных для анализа, выдачи отчетов и реализации других сервисных функций. В частности, для извлечения из БД информации, которая после оценки экспертом используется для обучения ЭВМ, и подготовки ее для последующего анализа разработана специальная подсистема. Она предоставляет эксперту возможности редактирования найденной информации и формирования выборки для анализа. В последнем случае эксперт только отмечает выбранные свойства компонентов в специальной таблице-меню и подсистема подготовки выборки для анализа извлекает выбранные значения свойств из БД “Elements”. Если нужно, в ней формируются алгебраические функции от исходных свойств и “склеивается” описание соединений в форме Excel-таблицы, которая затем поступает на вход прогнозирующей подсистемы. Подсистема выдачи результатов предназначена для предоставления

прогнозов в привычной для химиков и материаловедов табличной форме.

Важной особенностью программной реализации ИАС является то, что клиентская часть полностью построена на базе web-интерфейса [40]. Пользователи работают с ИАС посредством web-браузера. Процессы обучения и распознавания в ИАС реализованы с помощью специального асинхронного web-сервиса, что позволяет решать длительные по времени задачи обучения и прогнозирования в среде Интернет, в которой возможны сбои. Асинхронный web-сервис позволяет пользователям инициировать длительное выполнение ресурсоемких операций, контролировать степень их выполнения в асинхронном режиме, получать оповещение о готовых результатах расчетов, прерывать выполнение задач с сохранением промежуточных результатов.

ИСПОЛЬЗОВАНИЕ ИАС ДЛЯ ПРОГНОЗИРОВАНИЯ НОВЫХ СОЕДИНЕНИЙ И ОЦЕНКИ ИХ СВОЙСТВ

Результаты работы созданной ИАС проиллюстрированы таблицей, в которой приведены прогнозы возможности образования при нормальных условиях соединений состава AB_3X_3 в системах A_2X_3–B_2X (A и B – разные элементы; X = S, Se или Te) [44], перспективных для поиска новых полупроводниковых, нелинейно-оптических, электрооптических и акустооптических материалов.

Для компьютерного анализа использовалась экспериментальная информация о 117 примерах образования соединений состава AB_3X_3 и 58 примерах отсутствия соединений этого состава в системах A_2X_3–B_2X при нормальных условиях. Для описания соединений в памяти ЭВМ на основе физико-химических представлений о природе веществ этого типа были выбраны свойства элементов A, B и X (температуры плавления и кипения, ковалентный, ионный (по Бокию и Белову) и псевдопотенциальный (по Цангеру) радиусы, первые три потенциала ионизации, электроотрицательность (по Полингу), стандартные энтальпии атомизации и испарения, теплопроводности, молярные теплоемкости и т.д.), свойства простых халькогенидов A_2X_3 и B_2X (стандартная энтропия и энтальпия), а также некоторые алгебраические функции от этих свойств (например, отношение ковалентного радиуса к металлическому радиусу для элементов A, B и X [44]).

В таблице даны примеры прогнозов соединений состава AB_3X_3 и результаты их экспериментальной проверки. Приняты следующие обозначения: 1 – прогноз образования соединения AB_3X_3 при обычных условиях; 2 – прогноз отсутствия AB_3X_3 при обычных условиях; знаком # отмечены примеры, информация о которых использована для обучения ЭВМ; пустые клетки – неопределенный прогноз; © – совпадение прогноза образования соединения AB_3X_3 с новыми экспериментальными данными; Θ – совпадение прогноза отсутствия соединения с экспериментальными данными. Все 25 проверенных прогнозов совпали с экспериментальными данными [45–48].

Использование методов распознавания образов позволяет со средней точностью 80% прогнозировать, например, тип кристаллической структуры при заданных условиях [36, 37], оценивать некоторые фундаментальные свойства соединений, например, температуру плавления [49], ширину запрещенной зоны [30, 50] или критическую температуру перехода в сверхпроводящее состояние [51].

ЗАКЛЮЧЕНИЕ

Информационно-аналитическая система позволяет решить две важные задачи. Во-первых, она частично автоматизирует анализ огромной экспериментальной информации, накопленной в химической практике. Это позволяет найти закономерности в данных и применить их для конструирования новых соединений с заданными свойствами, причем использовать на этапе прогнозирования еще не полученных фаз только значения параметров компонентов. Во-вторых, ИАС расширяет возможности традиционных БД по свойствам веществ и материалов, предоставляя пользователю не только информацию об уже исследованных веществах, но и прогнозы для еще не изученных соединений с оценкой их свойств. Существенным преимуществом разработанной ИАС является ее доступность в сети интернет [52]. С помощью ИАС удалось получить прогноз тысяч еще не полученных неорганических соединений и оценить некоторые их свойства.

Работа выполнена при частичной финансовой поддержке РФФИ (проекты №№ 12-07-00142, 14-07-00819, 14-07-31032 и 15-07-00980).

СПИСОК ЛИТЕРАТУРЫ

1. *Belov G.V., Iorish V.S., Yungman V.S.* IVTANTHERMO for Windows – Database on Thermodynamic Properties and Related Software // CALPHAD: Comput. Coupling Phase Diagrams Thermochem. 1999. V. 23. № 2. P. 173.
2. *Bale C.W., Chartrand P., Degterov S.A. et al.* FactSage Thermochemical Software and Databases // CALPHAD: Comput. Coupling Phase Diagrams Thermochem. 2002. V. 26. № 2. P. 189.
3. *Yamashita Y., Yagi T., Baba T.* Development of Network Database System for Thermophysical Property Data of Thin Films // Jap. J. Appl. Phys. 2011. V. 50. № 11. P. 11RH03-1.
4. *Xu Y., Yamazaki M., Wang H., Yagi K.* Development of an Internet System for Composite Design and Thermo-

physical Property Prediction // Mater. Trans. 2006. V. 47. № 8. P. 1882.

5. *Yokokawa H., Yamauchi S., Matsumoto T.* Thermodynamic Database MALT for Windows with Gem and CHD // CALPHAD: Comput. Coupling Phase Diagrams Thermochem. 2002. V. 26. № 2. P. 155.
6. *Huang Z., Conway P.P., Thomson R.C. et al.* A Computational Interface for Thermodynamic Calculations Software MTDATA // CALPHAD: Comput. Coupling Phase Diagrams Thermochem. 2008. V. 32. № 1. P. 129.
7. *Andersson J.-O., Helander T., Hoglund L. et al.* THERMO-CALC & DICTRA, Computational Tools For-Materials Science // CALPHAD: Comput. Coupling Phase Diagrams Thermochem. 2002. V. 26. № 2. P. 273.
8. *Елецкий А.В., Еркимбаев А.О., Зицерман В.Ю., Кобзев Г.А., Трахтенгерц М.С.* Теплофизические свойства наноразмерных объектов: систематизация и оценка достоверности данных // ТВТ. 2012. Т. 50. № 4. С. 524.
9. *Фокин Л.Р., Калашников А.Н.* Транспортные свойства смеси разреженных газов N_2–H_2 в базе данных ЭПИДИФ // ТВТ. 2009. Т. 47. № 5. С. 675.
10. *Еркимбаев А.О., Зицерман В.Ю., Кобзев Г.А.* Систематизация данных по физико-химическим свойствам и применению углеродных наноструктур // ТВТ. 2010. Т. 48. № 6. С. 8694.
11. *Faber J., Fawcett T.* The Powder Diffraction File: Present and Future // Acta Crystallogr., Sect. B: Struct. Sci. 2002. V. 58. № 3. P. 325.
12. *Hellenbrandt M.* The Inorganic Crystal Structure Database (ICSD) – Present and Future // Crystallogr. Rev. 2004. V. 10. № 1. P. 17.
13. *White P.S., Rodgers J.R., Le Page Y.* CRYSTMET: a Database of the Structures and Powder Patterns of Metals and Intermetallics // Acta Crystallogr., Sect. B: Struct. Sci. 2002. V. B58. № 3. P. 343.
14. *Over H.H., Wolfart E., Dietz W., Toth L.* Mat-DB: A Web-Enabled Materials Database to Support European R&D Projects and Network Activities // Adv. Eng. Mater. 2005. V. 7. № 8. P. 766.
15. *Li X., Su H., Chen X. et al.* The Development of a Materials Database in China // Data Science J. 2007. V. 6. Suppl. P. S467.
16. *Киселева Н.Н., Дударев В.А., Земсков В.С.* Компьютерные информационные ресурсы неорганической химии и материаловедения // Успехи химии. 2010. Т. 79. № 2. С. 162.
17. http://iric.imet-db.ru.
18. *Киселева Н.Н., Дударев В.А.* База данных "Информационные ресурсы неорганической химии и материаловедения" // Информационные технологии. 2010. № 12. С. 63.
19. *Kimball R., Caserta J.* The Data Warehouse ETL Toolkit: Practical Techniques for Extracting, Cleaning, Conforming, and Delivering Data. N.Y.: John Wiley & Sons, 2004. 528 p.
20. *Morgenthal J.P.* Enterprise Information Integration: A Pragmatic Approach. Morrisville: Lulu. com, 2005. 317 p.
21. *Morgenthal J.P.* Enterprise Applications Integration with XML and Java. N.Y.–Sydney–London: Prentice Hall PTR, 2000. 528 p.
22. *Kornyushko V., Dudarev V.* Software Development for Distributed System of Russian DataBases on Electronics Materials // Int. J. Inform. Theor. Appl. 2006. V. 13. № 2. P. 119.
23. *Kiselyova N., Iwata S., Dudarev V. et al.* Principles of Integration of Russian and Japanese Databases on Inorganic Materials // Int. J. Inform. Technol. Knowledge". 2008. V. 2. № 4. P. 366.
24. *Киселева Н., Мурат Д., Столяренко А. и др.* База данных по свойствам тройных неорганических соединений "Фазы" в сети Интернет // Информационные ресурсы России. 2006. № 4. С. 21.
25. http://www.phases.imet-db.ru.
26. http://phases.imet-db.ru/elements.
27. *Христофоров Ю.И., Хорбенко В.В., Киселева Н.Н. и др.* База данных по фазовым диаграммам полупроводниковых систем с доступом из Интернет // Изв. вузов. Материалы электронной техники. 2001. № 4. С. 50.
28. *Киселева Н.Н., Прокошев И.В., Дударев В.А. и др.* Система баз данных по материалам для электроники в сети Интернет // Неорган. материалы. 2004. Т. 42. № 3. С. 380.
29. http://diag.imet-db.ru.
30. *Киселева Н.Н., Дударев В.А., Коржуев М.А.* База данных по ширине запрещенной зоны неорганических веществ и материалов // Материаловедение. 2015. № 7. С. 3.
31. http://www.bg.imet-db.ru.
32. http://crystal.imet-db.ru.
33. *Xu Y., Yamazaki M., Villars P.* Inorganic Materials Database for Exploring the Nature of Material // Jap. J. Appl. Phys. 2011. V. 50. № 11. P. 11RH02-1.
34. http://mits.nims.go.jp.
35. *Kiselyova N.N., Stolyarenko A.V., Ryazanov V.V. et al.* A System for Computer-Assisted Design of Inorganic Compounds Based on Computer Training // Pattern Recognition and Image Analysis. 2011. V. 21. № 1. P. 88.
36. *Kiselyova N., Stolyarenko A., Ryazanov V. et al.* Application of Machine Training Methods to Design of New Inorganic Compounds // Diagnostic Test Approaches to Machine Learning and Commonsense Reasoning Systems / Ed. Naidenova X.A., Ignatov D.I. Hershey: IGI Global. 2012. P. 197.
37. *Киселева Н.Н.* Компьютерное конструирование неорганических соединений. Использование баз данных и методов искусственного интеллекта. М.: Наука, 2005. 289 с.
38. *Журавлев Ю.И., Рязанов В.В., Сенько О.В.* "Распознавание". Математические методы. Программная система. Практические применения. М.: ФАЗИС, 2006. 176 с.
39. *Гладун В.П.* Процессы формирования новых знаний. София: СД "Педагог-6". 1995. 192 с.
40. *Столяренко А.В., Киселева Н.Н., Подбельский В.В.* Система компьютерного конструирования неорганических соединений // Автоматизация и современные технологии. 2008. № 9. С. 23.

41. *Senko O.V.* An Optimal Ensemble of Predictors in Convex Correcting Procedures // Pattern Recognition and Image Analysis. 2009. V. 19. № 3. P. 465.

42. *Yuan G.-X., Ho C.-H., Lin C.-J.* An Improved GLMNET for L_1-regularized Logistic Regression // J. Machine Learning Research. 2012. V. 13. P. 1999.

43. *Yang Y., Zou H.* A Coordinate Majorization Descent Algorithm for L_1 Penalized Learning // J. Statistical Computation & Simulation. 2014. V. 84. № 1. P. 1.

44. *Киселева Н.Н.* Прогнозирование существования AB_3X_3 (X = S, Se или Te) // Неорган. материалы. 2009. Т. 45. № 10. С. 1157.

45. *Schindler L.V., Schwarz M., Röhr C.* New Sulfido Antimonates of the Heavy Alkali Metals: Synthesis, Crystal Structure and Chemical Bonding of $(K/Rb/Cs)_3SbS_3$ and $Cs_3SbS_4 \cdot H_2O$ // Z. Naturforsch. B. 2013. V. 68. № 12. P. 1295.

46. *Олексеюк І., Цісар О., Піскач Л., Парасюк О.* Система $Tl_2Se–Ga_2Se_3$ // Наук. вісн. Східноєвр. нац. ун-ту ім. Лесі Українки. Серія: Хімічні науки. 2014. № 20. С. 26.

47. *Mucha I., Wiglusz K., Sztuba Z., Gawel W.* Phase Studies on the Quasi-Binary Thallium(I) Telluride-Gallium(III) Telluride System // Thermochim. Acta. 2011. V. 518. № 1–2. P. 53.

48. *Schwarz M., Röhr C.* Synthesis and Crystal Structure of Three New K-Thioferrates: K_9FeS_7, K_3FeS_3 and $K_{1.4}FeS_2$ // Z. Kristallogr. Suppl. 2012. V. 32. P. 96.

49. *Киселева Н.Н., Подбельский В.В., Рязанов В.В., Столяренко А.В.* Информационно-аналитическая система для конструирования новых неорганических соединений // Теплофизические свойства веществ и материалов. Тр. XII Рос. конф. по теплофиз. свойствам веществ. М.: Интерконтакт; Наука, 2009. С. 133.

50. *Kiselyova N.N., Stolyarenko A.V., Gu T. et al.* Computer-Aided Design of New Inorganic Compounds Promising for Search for Electronic Materials // Proc. 6th Int. Conf. Computer-Aided Design of Discrete Devices (CAD DD 07). V. 1. Minsk: UIPI NASB, 2007. P. 236.

51. *Савицкий Е.М., Киселева Н.Н.* Прогнозирование сверхпроводящих фаз Шевреля // Докл. АН СССР. 1978. Т. 239. № 2. С. 405.

52. http://ias.imet-db.ru.