СИСТЕМА КОМПЬЮТЕРНОГО КОНСТРУИРОВАНИЯ НЕОРГАНИЧЕСКИХ СОЕДИНЕНИЙ

А.В. Столяренко, Н.Н. Киселёва, В.В. Подбельский

Представлена информационно-аналитическая система для автоматизации процесса компьютерного конструирования новых неорганических соединений, основанная на использовании программ распознавания образов для поиска закономерностей в информации баз данных по свойствам неорганических веществ и материалов. Приведены результаты применения разработанной системы для прогнозирования новых соединений, перспективных для электроники.

Введение. Поиск и изучение новых веществ для использования в качестве активных компонентов твердотельной электроники является одной из важнейших задач химии и материаловедения. Появление уже первых ЭВМ обеспечило возможности для ускорения поиска новых неорганических веществ на основе компьютерного конструирования соединений-аналогов [1].

При конструировании ещё не полученных соединений необходимо найти совокупность химических элементов и их соотношение (т. е. качественный и количественный состав) для создания (при заданных внешних условиях) определенной пространственной молекулярной или кристаллической структуры соединения, позволяющей реализовать необходимые функциональные свойства. Эта задача может быть сведена к обнаружению зависимостей между свойствами физико-химических систем, в том числе свойствами соединений, и свойствами элементов, образующих эти системы.

Было показано [1], что методы обучения ЭВМ распознаванию образов являются одним из наиболее эффективных средств компьютерного конструирования неорганических соединений. Однако одной из трудностей, препятствующих широкому использованию систем распознавания образов в химической практике, является довольно сложная методика работы с этими системами. Если поиск информации в базах данных обычно является достаточно простой операцией, то подготовка найденной информации для её анализа с использованием программ распознавания образов требует от пользователя определённой квалификации. Самым перспективным путём решения этой проблемы является создание информационно-аналитической системы (ИАС), в которой автоматизированы не только поиск информации в базе данных (БД) по свойствам неорганических веществ и материалов (БД СНВМ), но и подготовка данных для анализа, отображение результатов прогнозирования в визуальной и табличной формах, поиск закономерностей в данных, а также хранение полученных закономерностей и прогнозов для дальнейшего использования. Именно такая информационно-аналитическая система для конструирования неорганических соединений, перспективных для применения в электронике, рассматривается в данной работе.

Этапы компьютерного конструирования с применением ИАС. Первый этап (1) компьютерного конструирования новых соединений (рис. 1) – это экспертный анализ информации баз данных по свойствам материалов для электроники и выбор соединений-прототипов. Соединение-прототип – это соединение с уже известными функциональными свойствами, которое либо уже используется, либо перспективно для использования в практической деятельности. Данные о свойствах этого соединения (или соединений) берутся из специализированных БД: по фазовым диаграммам полупроводниковых систем «Диаграмма» [1, 3], по свойствам акустооптических, электрооптических и нелинейнооптических веществ «Кристалл» [1, 4] и по ширине запрещённой зоны неорганических веществ «Bandgap» [1], в которых содержится информация о функционально важных для электроники свойствах веществ.

Пусть поставлена задача поиска новых полупроводниковых соединений. Анализ информации БД «Bandgap» и «Диаграмма» позволил выявить полупроводниковые соединения-прототипы состава BLi_3S_3, $TbCu_3S_3$, GaK_3Se_3.

Следующий этап (2) компьютерного конструирования – это выбор в БД по свойствам неорганических соединений «Фазы» [5] информации об отобранных на этапе (1) аналогах соединений-прототипов по составу – AB_3X_3 (A и B – здесь и далее различные химические элементы; X = S, Se, Te).

Соединение-аналог – это известное или неизвестное соединение, близкое по составу или кристаллической структуре соединению-прототипу. Сведения о соединениях-аналогах используются для «обучения» программы распознавания. Обучение ЭВМ – это процесс разделения объектов на альтернативные классы. Для обучения нужны данные о соединениях, аналогичных соединению-прототипу (например, по кристаллической структуре или составу) и альтернативных (например, системы, в которых соединение такого состава при определенных условиях вообще не образуется).

Для упомянутых выше соединений состава AB_3X_3 в БД «Фазы» запрашивается информация об известных системах с серой, селеном и теллуром, в которых образуются соединения прогнозируемого состава, и о системах, в которых при нормальных условиях такие халькогениды не обнаружены. Соответственно системы, в которых не образуются халькогениды состава AB_3X_3,

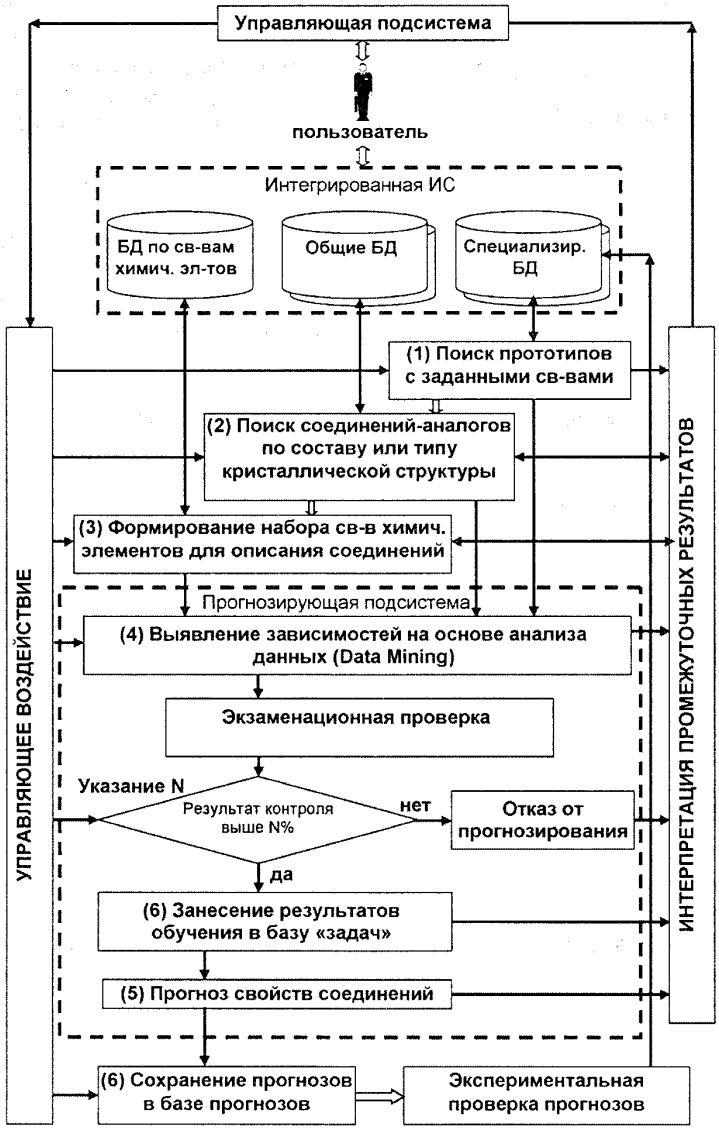

Рис. 1. Этапы компьютерного конструирования с применением информационно-аналитической системы

включены в альтернативный класс, а системы с образованием таких соединений – в целевой класс.

Каждое химическое соединение описывается в памяти ЭВМ в виде набора значений свойств химических элементов (3), входящих в его состав. Данные о свойствах элементов извлекаются из БД «Элементы». Результатом этого этапа является матрица, строки которой содержат описания систем в терминах свойств элементов и указания об их принадлежности к тому или иному классу систем (в нашем примере – к классам систем с образованием и без образования соединений состава AB_3X_3).

После предварительной обработки матрицы (например, удаления малоинформативных признаков и заполнения оставшихся в матрице пробелов) осуществляется процесс обучения (4).

На заключительном этапе (5) в найденную в результате обучения закономерность подставляются наборы значений свойств элементов – компонентов ещё неисследованных систем, и исследователь получает прогноз, будет ли образовываться в данной системе соединение заданного состава или нет.

Полученные закономерности и уже готовые прогнозы могут сохраняться в базе знаний (БЗ) для дальнейшего использования в ИАС (6).

Структура ИАС. В состав ИАС входят базы данных, подсистемы обучения и прогноза, база знаний и управляющая подсистема. Информационная основа компьютерного конструирования в ИАС – интегрированная распределённая система баз данных по свойствам неорганических веществ и материалов, созданная в ИМЕТ РАН [6]. Особенностью этой системы является объединение самых различных по программно-аппаратным платформам и качеству информации БД СНВМ.

Приведём кратко сведения об этих БД. Основным источником данных для компьютерного конструирования является БД по свойствам неорганических соединений «Фазы» [5]. В настоящее время она содержит информацию о свойствах около 44 тыс. тройных соединений (т. е. соединений, образованных тремя химическими элементами) из 18 тыс. систем (информация выбрана почти из 15 тыс. литературных источников). Сейчас в БД введена информация о более чем 15 тыс. четверных соединений.

БД «Диаграмма» [3] содержит собранную и оцененную высококвалифицированными экспертами информацию о фазовых P–T–x-диаграммах полупроводниковых систем и о физико-химических свойствах образующихся в них фаз. Сейчас в этой БД собраны данные о нескольких десятках систем.

БД «Кристалл» [4] включает информацию о более чем 100 веществах с особыми акустооптическими, электрооптическими и нелинейно-оптическими свойствами. Информация собрана и оценена квалифицированными экспертами в данной предметной области.

БД «Bandgap» [1, 6] создана для обеспечения специалистов информацией о ширине запрещённой зоны неорганических соединений. В настоящее время она содержит данные о более чем 3 тыс. веществ.

Необходимым элементом системы является БД «Элементы» по свойствам химических элементов. Она содержит данные о физических свойствах химических элементов, из которых конструируется описание химических соединений.

Прогнозирующая система, входящая в ИАС, основана на применении двух систем программ распознавания образов: «Распознавание» (российская фирма «Решения» совместно с Вычислительным центром им. А.А. Дородницына РАН) [7] и ConFor (Институт кибернетики им. В.М. Глушкова НАН Украины) [8].

Пакет программ «Распознавание» [7] содержит реализацию основных известных подходов в области распознавания образов и анализа данных (линейные, статистические, нейросетевые модели), а также новейшие разработки Российской академии наук (комбинаторно-логические модели, алгебраический подход, коллективные методы прогноза и кластеризации). Он включает программы на основе следующих методов и алгоритмов обучения ЭВМ: алгоритм вычисления оценок, алгоритм голосования по тупиковым тестам, алгоритм голосования по логическим закономерностям, алгоритм статистического взвешенного голосования, алгоритм линейной машины, алгоритм линейного дискриминанта Фишера, алгоритм k ближайших соседей, нейросетевой алгоритм распознавания образов с обратным распространением, алгоритм метода опорных векторов (support vector machine) и т. д. Используется также решение задач распознавания коллективами различных распознающих алгоритмов. Методы получения коллективных решений позволяют объединять исходные алгоритмы распознавания и получать некоторый новый алгоритм. Предполагается, что он будет сочетать в себе достоинства исходных методов и компенсировать недостатки каждого из них. Использование коллективных методов при решении большинства реальных задач позволяет повысить достоверность прогнозирования.

Система обучения ЭВМ формированию понятий ConFor [8] основана на оригинальном методе извлечения знаний на основе пирамидальных сетей.

Выбор вышеуказанных программных компонент анализа данных обусловлен:

универсальностью относительно размерностей решаемых задач. Системы способны решать как задачи прогноза редких или уникальных событий, явлений или процессов, когда начальная (обучающая) информация мала (десятки прецедентов), так и задачи больших размерностей (десятки тысяч прецедентов);

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

возможностью решения задач прогноза, классификации и анализа данных различными алгоритмами с последующим автоматическим построением оптимальных коллективных решений, что существенно повышает точность и надёжность прогноза.

В БЗ информационно-аналитической системы хранятся уже полученные закономерности для различных классов неорганических соединений. Эти закономерности могут использоваться для прогноза фаз и оценки их свойств, если в БД нет искомых сведений о конкретной химической системе. Для удобства пользователей БЗ дополнена базой прогнозов. В ней хранятся уже полученные прогнозы для различных классов соединений.

Управляющая система организует вычислительный процесс и осуществляет интерфейс между функциональными подсистемами ИАС, а так-

же обеспечивает доступ к системе из сети Интернет. Помимо этого, управляющая подсистема предоставляет пользователю программные средства подготовки данных для анализа, формирует отчёты в привычной для химиков форме, отображает результаты и реализует другие сервисные функции.

Программная реализация ИАС. Особенностью программной реализации ИАС является то, что клиентская часть полностью построена на базе Web-интерфейса. То есть пользователи работают с ИАС посредством Web-браузера.

Одним из важнейших требований, предъявляемых к ИАС, была возможность подключения разнообразных программ анализа данных с различными принципами работы, что позволяет улучшить качество прогнозов. При построении ИАС, объединяющей программы анализа данных, необходимо обеспечить хранение информации об этих программах и методах распознавания образов, реализованных в них. Для хранения данных было принято решение использовать реляционную базу метаданных, именуемую в дальнейшем метабазой. Для решения задачи объединения программ анализа данных предлагается следующая структура метабазы, состоящая из двух частей. Первая группа таблиц отвечает за хранение информации непосредственно об интегрируемых программах, предоставляемых ими функциях и способах вызова этих функций. Вторая группа таблиц отвечает за хранение информации о настраиваемых параметрах методов обучения.

Взаимодействие между программами анализа данных, которые реализуют методы обучения и распознавания, и управляющей подсистемой происходит посредством программных адаптеров, предоставляющих все необходимые функции программы, что соответствует идеям «сервисно-ориентированного» подхода. Для интеграции новой программы анализа данных в ИАС нужен только программный адаптер, выполняющий сопряжение внутренних структур данных интегрируемой информационной системы со стандартизированным представлением данных в интегрированной системе.

Адаптер интегрируемой программы анализа данных предоставляет следующие средства: обучение соответствующего метода анализа данных с заданными параметрами, экзамен на обучающей выборке, распознавание с использованием ранее обученного метода. Информация о функциях и их аргументах хранится в метабазе, содержащей данные об интегрируемых программах.

Разработана архитектура системы, которая позволяет пользователям инициировать длительное по времени выполнение ресурсоёмких операций, контролировать степень их выполнения в асинхронном режиме и получать оповещение о готовых результатах расчётов.

При программной реализации БЗ возникла проблема, связанная с тем, что форма представления знаний в используемых методах распознавания образов существенно различается. В связи с этим было предложено новое программное решение для хранения полученных закономерностей, а также сопутствующей информации о параметрах программ и исследуемых объектов. 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 признаков, используемых для описания объектов, пути к файлам на дисках, фамилия специалиста, проводившего оценку данных для обучения и поиск закономерностей, дата получения закономерности и т. д.

Результаты использования ИАС для конструирования неорганических соединений, перспективных для электроники. С помощью разработанной ИАС решаются следующие задачи конструирования химических систем:

с образованием и отсутствием соединений;

с образованием соединений с определённым соотношением компонентов;

с заданным типом кристаллической структуры;

с заданными функциональными свойствами (температурой плавления, температурой перехода в сверхпроводящее состояние, шириной запрещенной зоны и т. д.).

Таким образом, речь идёт о нахождении (среди возможных комбинаций различных элементов) аналогов уже известных соединений, обладающих искомыми свойствами.

В табл. 1 приведены результаты прогнозирования образования соединений состава AB_3S_3 (применён метод «линейная машина»). По горизонтали расположены элементы «B», а по вертикали элементы «A» из формулы соединения AB_3S_3. В ячейке таблицы на пересечении столбца и строки выводится прогноз для соответствующего химического соединения. После символа «#» указывается класс объекта из обучающей выборки.

В табл. 2 приведена точность прогноза образования сульфидов AB_3X_3 с использованием различных алгоритмов распознавания образов, включенных в ИАС. Экзаменационное распознавание проводилось на изначально неиспользованной для обучения части массива обучающей выборки, после чего проводилось переобучение по всей обу-

Таблица 1

Пример результатов прогнозирования (1 – соединение образуется, 2 – не образуется, # – объекты обучающей выборки)

A / B	Cr	Mn	Fe	Co	Ni	Ga	As	Y	In	Sn	Sb	La	Pm	Sm	Eu	Gd	Tb	Er	Tm	Yb	Lu	Ir	Au	Tl	Bi
Li	1	2	2	2	1	2	1	1	2	1	1	2	2	2	2	1	2	2	1	1	1	1	1	1	2
Na	1	1	#1	1	1	#1	#1	1	1	1	#2	2	1	1	1	1	1	1	1	1	1	1	1	1	2
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
Cu	1	1	1	1	1	#2	#1	#1	#1	#1	#1	2	2	#1	#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
Ag	2	2	2	2	2	#2	#1	2	2	1	#2	2	2	2	2	1	2	2	2	2	2	1	1	#2	2
In	1	2	2	2	2	2	#1	2		1	2	2	2	2	2	1	2	2	2	2	2	1	1	2	2
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
Au	2	2	2	2	2	2	1	2	2	2	2	2	2	2	2	2	2	2	2	2	2	1		2	#2
Tl	#1	2	2	2	1	2	#1	1	#2	#1	#1	2	2	2	2	1	2	2	1	1	1	1	1		2

Таблица 2

Точность прогноза образования халькогенидов AB_3X_3 различными методами

Алгоритм	Достоверность, %
Алгоритмы вычисления оценок	74,8
Линейный дискриминант Фишера	75,4
Линейная машина	80,5
Логические закономерности	81,8
Многослойный перцептрон	81,0
Q ближайших соседей	82,3
Метод опорных векторов	84,4

чающей выборке. Следует отметить, что точность прогноза достаточно высокая – выше 76 %, что свидетельствует о возможности правильной оценки образования халькогенидов состава AB_3X_3.

ИАС снабжена удобными средствами визуализации, позволяющими отображать любую проекцию или сечение многомерного пространства свойств компонентов. На рис. 2 показана диаграмма распределения проекций точек, соответствующих перспективным для поиска новых пьезоэлектрических материалов соединениям состава ABO_3 с различными типами кристаллической структуры, на плоскость с координатами ($2A+3B$; A/B), где A и B – первые энергии ионизации химического элемента A и B соответственно.

Разработанная система визуализации, с одной стороны, позволяет провести отбор свойств химических элементов, наиболее важных для классификации образуемых ими соединений, а с другой стороны, является удобным средством графической иллюстрации выявленных закономерностей.

Заключение. Новизна описанной ИАС состоит в том, что впервые в данной предметной области проведена интеграция в единую систему достаточно разнородных информационных компонентов. В ИАС используются математические методы и программные средства синтеза признаковых описаний с заданными прогнозными свойствами на базе основных моделей распозна-

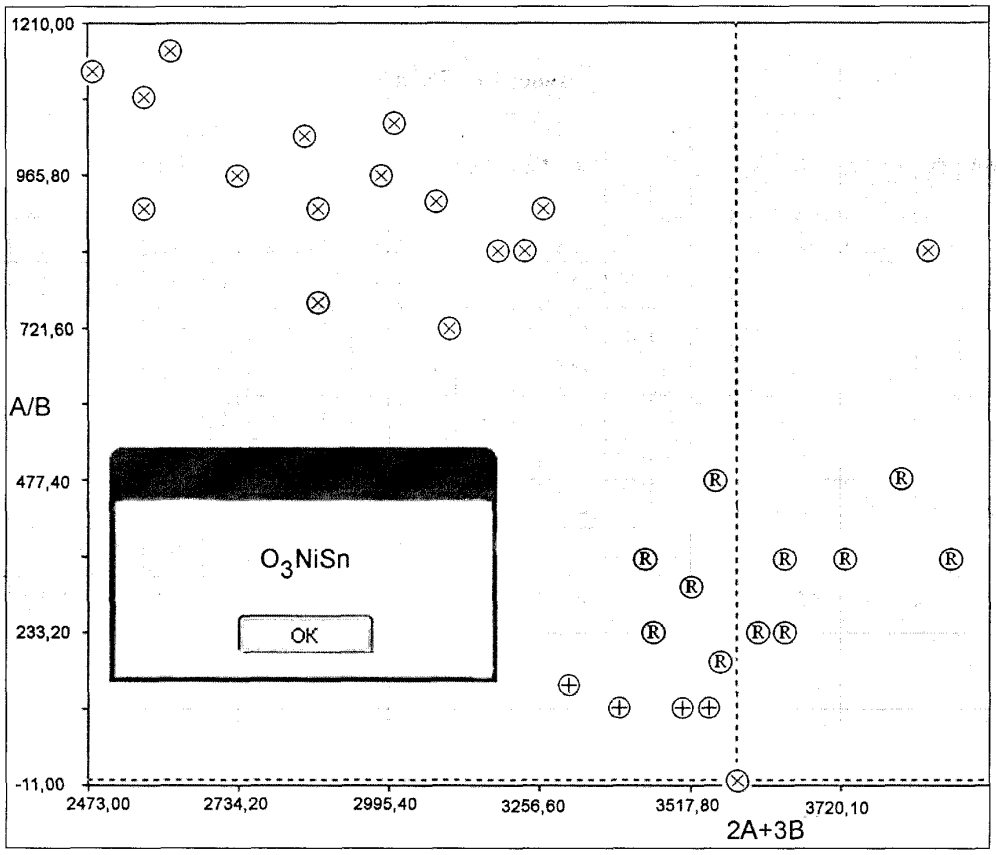

Рис. 2. Диаграмма распределения соединений:
⊗ - перовскит; ® - ильменит; ⊕ - волластонит

вания и интегрированная система баз данных по свойствам неорганических соединений. Применение ИАС позволяет значительно увеличить достоверность прогноза новых неорганических соединений за счёт использования больших выборок для обучения ЭВМ, выбора «хорошего» набора признаков для описания химических систем и принятия решения о принадлежности прогнозируемых неорганических соединений к тому или иному классу фаз на основе сравнения результатов прогноза, полученных с применением различных программ обучения ЭВМ. Работа выполнена при поддержке РФФИ (грант № 06-07-89120).

Список литературы

1. **Киселёва Н.Н.** Компьютерное конструирование неорганических соединений. Использование баз данных и методов искусственного интеллекта. М.: Наука, 2005.

2. **Христофоров Ю.И., Хорбенко В.В., Киселева Н.Н. и др.** База данных по фазовым диаграммам полупроводниковых систем с доступом из Интернета. *Изв. ВУЗов. Материалы электронной техники.* 2001. № 4.

3. **Киселёва Н.Н., Прокошев И.В., Дударев В.А. и др.** Система баз данных по материалам для электроники в сети Интернет. *Неорганические материалы.* 2004. Т. 42. № 3.

4. **Киселёва Н.Н., Подбельский В.В., Столяренко А.В. и др.** База данных по свойствам тройных неорганических соединений «Фазы» в сети Интернет как основа компьютерного конструирования новых материалов. *Информационные ресурсы России.* 2006. № 4.

5. **Дударев В.А., Киселёва Н.Н., Земсков В.С.** Интегрированная система баз данных по свойствам материалов для электроники. *Перспективные материалы.* 2006. -№5.

6. **Журавлев Ю.И., Рязанов В.В., Сенько О.В.** «Распознавание». Математические методы. Программная система. Практические применения. М.: Фазис, 2006.

7. **Гладун В.П.** Процессы формирования новых знаний. София: СД «Педагог-6», 1995.