МЕХАНИЗМЫ ИНТЕГРАЦИИ БАЗ ДАННЫХ И ПРОГРАММ АНАЛИЗА

А.В. Столяренко,
кандидат технических наук, научный сотрудник Учреждения Российской академии наук Института металлургии и материаловедения им. А.А. Байкова РАН, e-mail: stol-drew@yandex.ru.
Н.Н. Киселева,
доктор химических наук, заведующая лабораторией Учреждения Российской академии наук Института металлургии и материаловедения им. А.А. Байкова РАН, e-mail: kis@imet.ac.ru.
В.В. Подбельский,
доктор технических наук, профессор Государственного университета – Высшей школы экономики, e-mail: vpodbelskiy@hse.ru.
Адрес: г. Москва, Ленинский проспект, 49.

В статье предложен подход к интеграции баз данных и программ анализа данных. Описано его применение при разработке информационно-аналитической системы для автоматизации процесса компьютерного конструирования неорганических соединений. Рассматриваемые принципы интеграции разнородных программных и информационных компонентов могут быть применены и в других предметных областях: в медицине, экономике, промышленности, бизнесе.

Ключевые слова: база данных, информационно-аналитическая система, анализ данных, интеграция разнородных программных и информационных систем, распознавание образов, сервисно-ориентированный подход, метабаза.

1. Введение

Появление многочисленных баз данных в различных предметных областях поставило перед специалистами вопрос рационального использования хранящейся в них информации не только для информационного обслуживания, но и для анализа с целью выявления зависимостей в данных и прогнозирования неизвестных значений параметров объектов. Одно из наиболее перспективных и актуальных направлений связано с разработкой информационно-аналитических систем (ИАС), объединяющих базы данных и программы анализа данных. ИАС автоматизирует хранение и изменение информации, подготовку данных для анализа, проведение прогнозирования, визуализацию и отображение результатов анализа данных. С помощью таких систем, в частности, удается найти взаимосвязи между различными объектами и выявлять закономерности, присущие предметной области информационной системы.

2. Постановка задачи

ИАС зачастую создаются на основе уже существующих баз данных, информационных систем и программ анализа данных. В связи с этим актуальной задачей является интеграция разнородных программных и информационных компонентов. Решение проблемы усложняется, если информационным источником в ИАС является система баз данных, созданных в разное время и на основе различных систем управления базами данных (СУБД). Кроме того, в ИАС, в общем случае необходимо включать разные по идеологии средства обработки данных. Перспективным является проведение интеллектуального анализа данных с применением программных решений не только в локальной среде, но и в сети Интранет и Интернет.

3. Информационно-аналитическая система для компьютерного конструирования неорганических соединений

В настоящей работе рассмотрены принципы разработки ИАС для информационного обслуживания специалистов в области неорганической химии и материаловедения [1]. ИАС предназначена для конструирования новых неорганических соединений с заданными свойствами. Ее применение дает возможность найти сложные зависимости между фундаментальными свойствами неорганических соединений и фундаментальными свойствами химических элементов. Использование найденных взаимосвязей позволяет проводить компьютерное конструирование неорганических соединений [1, 2] и оценивать различные их свойства без реального синтеза этих соединений.

В состав ИАС входят программы анализа данных, подсистема визуализации результатов, база полученных закономерностей и прогнозов и управляющая подсистема. Управляющая подсистема организует вычислительный процесс и осуществляет взаимодействие между функциональными подсистемами ИАС, а также обеспечивает доступ к системе из сети Интернет. Помимо этого, управляющая подсистема предоставляет пользователю программные средства подготовки данных для анализа, выдачи отчетов в привычной для химиков форме, визуализации результатов и реализации других сервисных функций.

Существуют две задачи, требующие решения при разработке ИАС: задача интеграции баз данных и задача интеграции программ анализа данных.

При разработке принципов интеграции баз данных по свойствам неорганических веществ и материалов [2-7] принималась во внимание специфика предметной области: базы данных распределены по различным организациям-разработчикам, в них хранится информация с разным уровнем достоверности, использованы различные операционные системы, форматы данных и СУБД. В связи с вышеуказанной спецификой предметной области, общепринятые методики объединения информационных систем (ИС) на основе хранилища данных оказались непригодны. Применен комплексный подход к интеграции, сочетающий в себе интеграцию на уровне данных и пользовательских интерфейсов [5-7]. В рамках предлагаемого подхода пользователю предоставляется доступ к текущим пользовательским интерфейсам баз данных, свободное перемещение между ними, и богатые возможности по агрегации информации, полученной из разнородных распределенных источников данных по свойствам веществ, согласно общей разработанной информационной схеме.

Необходимость интеграции программ была обусловлена тем, что для улучшения качества прогнозирования в ИАС используются специальные коллективные методы принятия решения [8], в процессе функционирования которых взаимодействуют программы анализа данных [8, 9] с различными принципами работы. При этом решение задачи интеграции программ должно быть:

- ✦ масштабируемым, т.е. обеспечивать возможность поэтапного добавления программ анализа данных в ИАС;
- ✦ достаточно простым для реализации, чтобы разработка программных модулей для включения новой программы анализа данных в ИАС на основе предложенной методики не представляла сложной задачи;
- ✦ гибким, чтобы учитывать различия в данных и информационных структурах программ;
- ✦ мощным, чтобы обеспечить сложные механизмы взаимодействия программ анализа данных.

При решении задачи интеграции был применен сервисно-ориентированный подход (SOA). SOA — прикладная архитектура, в которой все функции определены как независимые сервисы с четкими интерфейсами [10]. Обращение к этим сервисам в определенной последовательности позволяет реализовать тот или иной процесс.

Взаимодействие между подсистемами анализа данных, которые реализуют все методы обучения и распознавания, и управляющей подсистемой происходит посредством программных адаптеров, предоставляющих все необходимые функции программы анализа данных, что соответствует идеям SOA. 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.

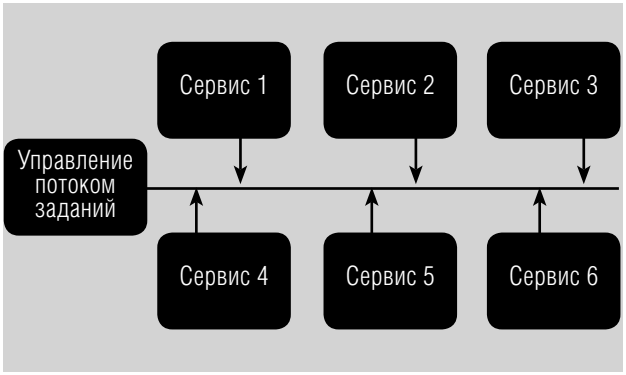

Рис. 1. Модель сервисной шины.

Следует обратить внимание на архитектурную концепцию, используемую для сервисно-ориентированной интеграции. Речь идет о концепции сервисной шины. Ее задача — предоставить единый механизм передачи запросов и получения результатов сервисов, выполнения необходимых преобразований сообщений и транспортных протоколов, и, что наиболее важно, управления потоком обращений к сервисам. Благодаря такому управлению, упрощается организация нужной последовательности вызовов сервисов для реализации процесса. Обратившись к схематичной иллюстрации шины *(рис. 1)*, можно увидеть, что этот подход решает одну из главных проблем интеграции — проблему минимизации интерфейсов.

Заметим, что независимо от выбранной технологии интеграции модулей всегда требуется разрабатывать специальные программы-«адаптеры» для каждой функции каждого приложения, обеспечивающие выгрузку или загрузку передаваемых данных. Эти «адаптеры» оперируют внутренним представлением данных конкретного приложения. Для обеспечения технологичного взаимодействия «адаптеров» приложений целесообразно принять «межмодульный» формат представления данных. В настоящее время наиболее удобным средством для описания «межмодульного» формата является язык XML [11]. При его использовании необходимо определить «Пространство имен XML-документов». Пространство имен – это коллекция имен, используемых в XML-документах в качестве атрибутов и элементов, поименованная с помощью унифицированного идентификатора ресурсов – словаря разметки. Прикладные XML-форматы, использующие конкретное «Пространство имен», являются в конечном итоге «межмодульными» форматами представления данных.

При программной реализации ИАС используется архитектура «клиент-сервер». Вся вычислительная работа происходит на Web-сервере, а пользователю выдаются только результаты для просмотра. Такая организация позволяет легко расширять ИАС, интегрируя в нее новые методы анализа данных, добавлять различную функциональность без необходимости обновления клиентского приложения. За счет оснащения ИАС Web-интерфейсом пользователи могут проводить анализ данных через Интернет.

Адаптер интегрируемой в ИАС программы анализа данных должен предоставлять следующие средства: обучение с использованием соответствующего метода анализа данных с заданными параметрами, экзамен на обучающей выборке, распознавание с использованием ранее примененного метода обучения ЭВМ. Информация об этих средствах, реализованных в виде функций, и о параметрах этих функций хранится в метабазе – справочной базе метаданных, содержащей информацию об интегрируемых программах анализа *(рис. 2)*.

Таблица метабазы *(рис. 2)* MetaPrograms является главной таблицей со списком программ, подключенных к ИАС. Каждой подключаемой программе присваивается уникальный целочисленный идентификатор ProgramID. В данной таблице также содержится информация, необходимая для сопряжения программ с ИАС. Поле Name – название программы. В поле PathWrappper указан путь к адаптеру программы. В поле ProjectsDir хранится адрес каталога сервера, в котором система сохраняет результаты своей работы.

В таблице MetaFunctionTypes содержится информация о типах функций программ анализа данных, а именно: обучение, обучение коллективным методом, распознавание.

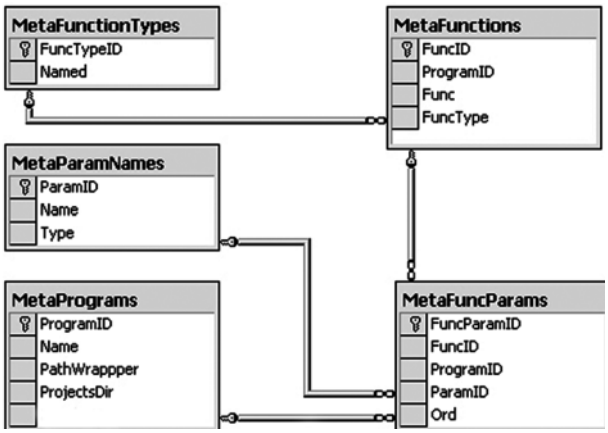

Рис. 2. Структура таблиц метабазы.

В таблице MetaFunctions содержится информация о функциях программ анализа данных, доступных «извне».

В таблице MetaParamNames содержится информация о параметрах функций. Поле Name содержит имя параметра. Используются следующие параметры: путь к текущему рабочему каталогу, главный метод обучения, список методов (используется в случае коллективного решения [8]), количество методов, обучающая выборка, обучающая выборка для коллективного решения, идентификатор функции, список параметров методов распознавания.

В таблице MetaFuncParams содержится информация о параметрах конкретных функций программы анализа.

При реализации ИАС возникают следующие проблемы. Работа некоторых методов обучения может продолжаться несколько часов (особенно для больших выборок). Естественно пользователь ИАС не должен все это время поддерживать связь с системой. Очевидно и то, что может произойти сбой в сети Интернет. Для того, чтобы избежать обозначенных проблем, удобно работу процессов обучения и распознавания реализовать с помощью асинхронного Web-сервиса [12]. Такой подход позволяет реализовать механизм сохранения инициированных пользователем процессов, а также предоставлять ему информацию о ходе выполнения той или иной операции. При повторном входе в систему пользователь получает текущее состояние инициированных им процессов.

Рассмотрим варианты построения такого асинхронного Web-сервиса.

Известно [13], что клиентский прокси-класс, генерируемый для обращения к Web-сервису, содержит как синхронный вариант вызова методов сервиса, так и асинхронный. Если, например, в Web-сервисе используется метод Method, то в прокси-классе будет сгенерирован соответствующий синхронный метод Method и пара методов для асинхронного вызова – BeginMethod и EndMethod.

Очевидно, что синхронный вариант для запуска длительных серверных процессов не подходит, так

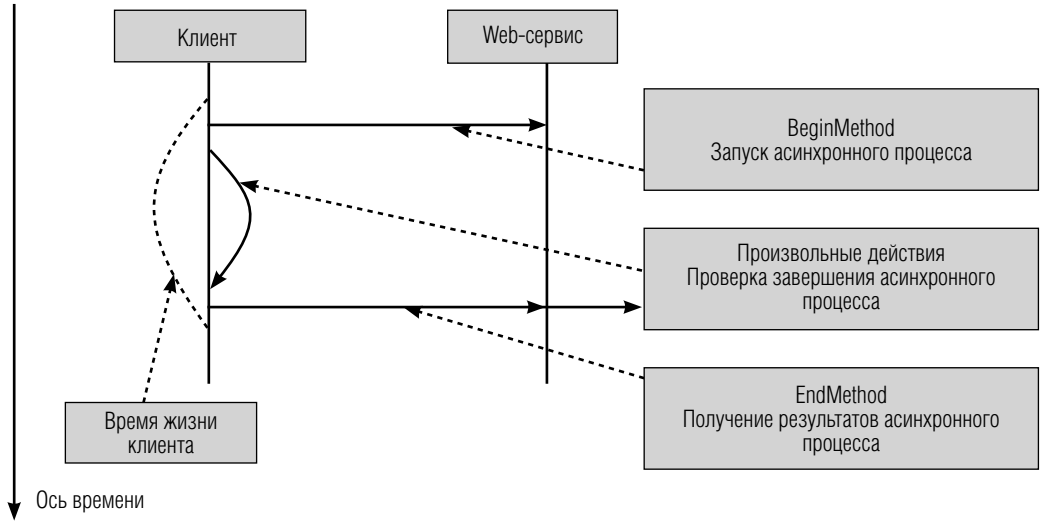

Рис. 3. Схема стандартного асинхронного использования Web-методов.

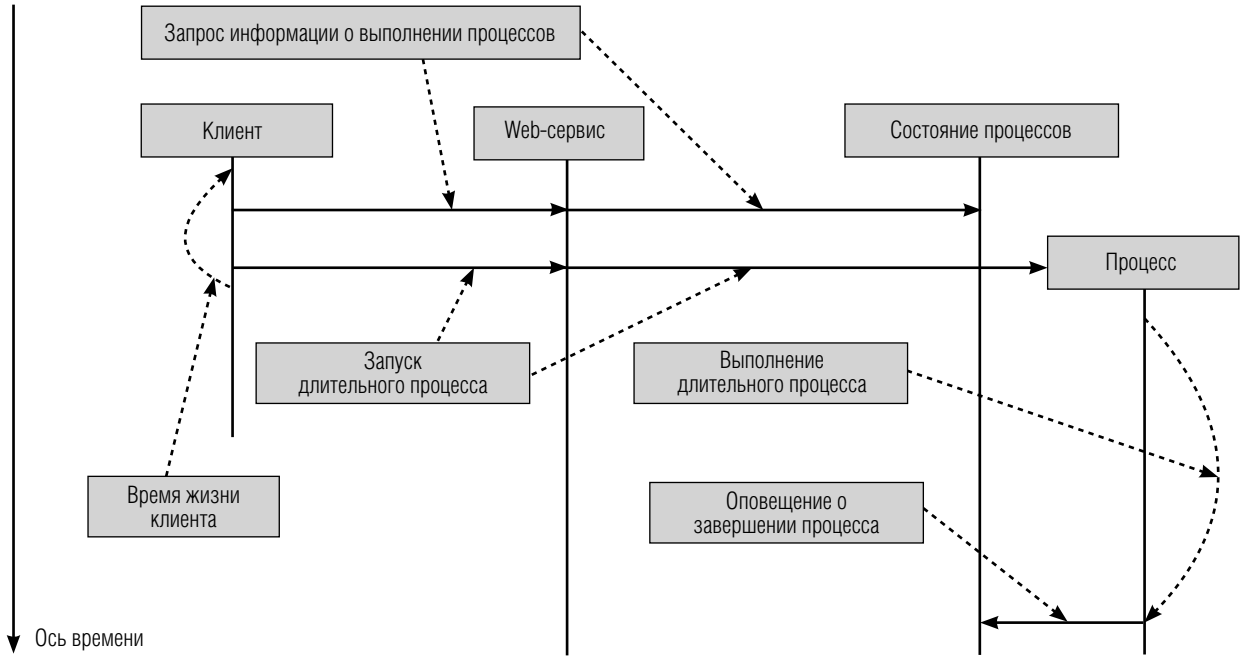

Рис. 4. Модель асинхронного Web-сервиса.

как время реакции системы, исчисляемое десятками минут, просто неприемлемо.

Предлагаемая разработчиками Microsoft схема асинхронного использования Web-методов [12] может быть представлена в виде диаграммы *(см. рис. 3)*.

Время отклика системы при такой схеме работы в ряде случаев окажется непозволительно большим.

В настоящей работе предложена архитектура системы *(рис. 4)*, которая позволяет пользователям инициировать длительное выполнение ресурсоемких операций, контролировать степень их выполнения в асинхронном режиме и получать оповещение о готовых результатах расчетов. Здесь «Состояние Процессов» – некоторый глобальный (в контексте Web-сервиса) объект, который хранит информацию о выполнении процессов, а «Процесс» – объект, непосредственно выполняющий назначенный процесс. «Процесс» имеет возможность записывать в объект «Состояние Процессов» информацию о своем выполнении. Таким образом, используются три компонента: класс Web-сервиса, класс состояния процессов и непосредственно класс, реализующий сам процесс.

Остановимся подробнее на методах этих классов.

Класс «Процесс» отвечает за выполнение текущей операции и содержит методы «Старт», «Остановка» и «Информация о состоянии». Метод «Старт» осуществляет запуск нового процесса, сохраняет время запуска и идентификатор процесса в переменных класса. Метод «Остановка» позволяет остановить ранее запущенный процесс. Метод «Информация о состоянии» возвращает информацию о состоянии запущенного процесса в виде структуры, содержащей время работы процесса и флаг, показывающий состояние: «выполняется», «завершен корректно», «завершен с ошибкой».

Класс «Состояние Процессов» выполняет функции промежуточного слоя между Web-сервисом и классом «Процесс». Для хранения множества запросов используется хэш-таблица и таблица в базе данных. Класс содержит методы «Старт», «Остановка» и «Информация о состоянии». Метод «Старт» создает объект класса «Процесс» и инициирует с его помощью необходимую операцию. Методы «Остановка» и «Информация о состоянии» по соответствующему идентификатору останавливают или получают информацию о ходе запущенного процесса.

Внешним уровнем модели асинхронного сервиса является сам Web-сервис. В его функции входит не только пересылка запросов пользовательскому компоненту, но и выполнение таких операций, как создание, размещение и сохранение экземпляра компонента, который может использоваться все-

ми клиентами. Web-сервис содержит метод «Запуск процесса», позволяющий инициировать процесс. Метод «Запуск процесса» использует набор аргументов для запуска процесса, который передается с помощью объекта класса «Аргументы службы». Класс «Аргументы службы» содержит всю необходимую информацию для запуска процесса: используемые методы обучения, их параметры, выборка для обучения или прогнозирования в формате XML, или путь к файлу с уже сохраненной на сервере выборкой.

Клиентом разработанного Web-сервиса может быть как серверный код ASP.NET-страницы, так и Windows-приложение, что обеспечивает достаточно гибкую реализацию запуска длительных задач.

При реализации ИАС важен выбор единого формата выборок для обучения и прогнозирования. Его соблюдение облегчает подключение к ИАС новых программ анализа данных и взаимодействие между ними.

Выборка для обучения готовится средствами ИАС в формате XML следующей структуры:

```
<Selection NumProperties="Количество признаков"
NumObjects="Количество объектов">
<Object name="Название объекта"
Class="Классообразующий признак">
<Property name="Название признака"
value="Значение признака"/>
<Property name="Название признака "
value="Значение признака " />
...
</Object>
<Object name="Название объекта"
Class="Классообразующий признак">
...
</Object>
</Selection>
```

По результатам обучения, экзамена и прогнозирования формируются отчеты в формате HTML, доступные для просмотра, а результаты распознавания представляются в виде XML. Они имеют следующий формат:

```
<Prediction>
<Object name="Название объекта"
PredictedClass="Имя класса">
</Object>
<Object name="Название объекта "
PredictedClass="Имя класса">
...
</Object>
...
</Prediction>
```

Для каждого объекта указывается его принадлежность к тому или иному классу объектов.

Выявленные экспертом в результате работы с ИАС закономерности сохраняются во внутреннем формате программы анализа данных, с помощью которой они были получены. При этом в базе «задач» сохраняются не сами закономерности (например, логические выражения или структура обученной нейронной сети), а так называемые «ярлыки» для этих «задач». Под термином «ярлык» понимается вся необходимая информация о «задаче», позволяющая идентифицировать ее среди остальных: идентификатор программы анализа данных, с помощью которой производилось обучение; путь к файлам на сервере; список методов и их параметры; признаки, использованные при формировании выборок; изучаемая характеристика объектов; а также сведения о количественном и качественном составе химических соединений, информация о которых использовалась для обучения. Такая реализация позволяет достаточно просто встраивать в ИАС новые программы анализа данных и решает проблему, связанную с тем, что форма представления знаний в используемых методах обучения ЭВМ существенно различается.

4. Заключение

Предложенные подходы и алгоритмы применены для создания информационно-аналитической системы для компьютерного конструирования неорганических соединений. Разработанная ИАС была успешно применена для компьютерного конструирования новых халькогенидных соединений, перспективных для использования в качестве полупроводниковых и магнитных материалов [15-17]. Экспериментальная проверка результатов, полученных с помощью ИАС, показала, что точность прогноза новых соединений выше 80 %.

Работа выполнена при поддержке РФФИ (гранты №06-07-89120, 08-01-90427, 08-07-00437, 05-03-39009 и 09-07-00194).

Авторы выражают благодарность В.А. Дудареву за ценные замечания. ■

5. Литература

1. Information-analytical system for design of new inorganic compounds / N. Kiselyova, A. Stolyarenko, V. Ryazanov, V. Podbel'skii // Int.J. «Information Theories & Applications». 2008. Vol. 2. N. 4. P. 345-350.
2. Киселева Н. Н. Компьютерное конструирование неорганических соединений. Использование баз данных и методов искусственного интеллекта. М. : Наука, 2005. – 288 с.
3. База данных по свойствам по свойствам тройных неорганических соединений «Фазы» в сети Интернет как основа компьютерного конструирования новых материалов. / Н. Киселева, Д. Мурат, А. Столяренко, В. Дударев, В. Подбельский, В. Земсков // Информационные ресурсы России. – 2006. – N. 4. – С. 21 – 23.
4. База данных по фазовым диаграммам полупроводниковых систем с доступом из Интернет / Ю.И. Христофоров, В.В. Хорбенко, Н.Н. Киселева, В.В. Подбельский, И.Н. Белокурова, В.С. Земсков // Изв.ВУЗов. Материалы электронной техники. – 2001. – № 4. – С. 50 – 55.
5. Система баз данных по материалам для электроники в сети Интернет / Н.Н. Киселева, И.В.Прокошев, В.А. Дударев, В.В. Хорбенко, И.Н. Белокурова, В.В.Подбельский, В.С. Земсков // Неорганические материалы. – 2004. – Т. 42, № 3. – С. 380 – 384.
6. Integration principles of Russian and Japanese databases on inorganic materials / N.Kiselyova, S. Iwata, V. Dudarev, I. Prokoshev, V. Khorbenko, V. Zemskov // Int.J. "Information Technologies and Knowledge". 2008. Vol. 2, № 4. P. 366-372.
7. Киселева Н. Н., Дударев В. А., Земсков В. С. Компьютерные информационные ресурсы неорганической химии и материаловедения // Успехи химии. – 2010. – Т. 79, № 2. – С. 162-188.
8. Журавлев Ю. И., Рязанов В. В., Сенько О. В. «Распознавание». Математические методы. Программная система. Практические применения. М. : ФАЗИС, 2006. – 176 с.
9. Гладун В. П. Процессы формирования новых знаний. София : СД «Педагог 6», 1995. – 192 с.
10. Migrating to a service-oriented architecture. / K. Channabasavaiah, K. Holley, E.M. Tuggle. // IBM, December 2003.
11. http://www.w3.org/XML/ (дата обращения: 31.05.2010).
12. Ньюкомер Э. Веб-сервисы: XML, WSDL, SOAP и UDDI. С.Петербург : Изд.: Питер, 2003. – 256 с.
13. http://msdn.microsoft.com/library/rus/vbcon/html/vbtskCallingWebServiceAsynchronously.asp (дата обращения: 31.05.2010).
14. http://msdn.microsoft.com/library/rus/cpref/html/frlrfsystemwebhttpapplicationstateclasstopic.asp (дата обращения: 31.05.2010).
15. Киселева Н. Н. Прогнозирование существования AB_3X_3 (X = S, Se,Te) // Неорганические материалы. – 2009. – Т. 45, № 10. – С. 1157-1160.
16. Бурханов Г. С., Киселева Н. Н. Прогнозирование интерметаллических соединений // Успехи химии. – 2009. – Т. 78, № 6. – С. 615-634.
17. Компьютерное конструирование новых неорганических соединений состава ABX2 (X = S, Se или Te / Н.Н. Киселева, В.В. Подбельский, В.В. Рязанов, А.В. Столяренко // Материаловедение. – 2008. – № 12. – С. 34-41.