Организация и использование информационных ресурсов

Туманов Владимир Евгеньевич *- кандидат химических наук, заведующий лабораторией информационной поддержки научных исследований Института проблем химической физики (ИПХФ) РАН*
e-mail: tve@icp.ac.ru
Прохоров Андрей Иванович *- научный сотрудник ИПХФ РАН*
e-mail: aipro@icp.ac.ru
Лазарев Даниил Юрьевич *- инженер-исследователь ИПХФ РАН*
e-mail: tve@icp.ac.ru
Соловьева Маина Ефимовна *- научный сотрудник ИПХФ РАН*
e-mail: solov@icp.ac.ru

Разработка и создание предметно-ориентированной системы научной осведомленности по физической химии радикальных реакций

Введение

Успешное использование и развитие систем деловой осведомленности или бизнес-аналитики (Business Intelligence System, BI System) привело к идее разработки и создания систем научной осведомленности (Science Intelligence System, SI System). В статье [1] ***система научной осведомленности*** определена «как информационная инфраструктура, которая обеспечивает принятие решений и совместную работу научного сообщества в рамках выделенной предметной области знаний». Там же рассмотрена общая архитектура таких систем в разрезе категорий основных пользователей и использования современных информационных технологий.

В основу создания таких систем положены технологии складирования данных (Data Warehousing), анализа и извлечения знаний (Data Mining) [2]. В системах научной осведомленности научные решения, методология и методы исследований интегрируются в общую библиотеку решений, а данные из разнородных источников интегрируются в общее хранилище данных, которое через предметно-ориентированные информационные ресурсы поставляет информацию пользователям: ученым, технологам, представителям промышленности и государства. Особенностью систем научной осведомленности является предоставление пользователям, помимо собственно профессиональной информации, инструментария для анализа данных.

В докладе [3] был рассмотрен общий подход к построению систем управления фундаментальными знаниями на примере физико-химических данных. В докладе [4] был использован аналогичный подход для создания интеллектуальной информационной системы по физической химии радикальных реакций, при этом модель системы рассматривалась с точки зрения внедрения в нее элементов прикладного искусственного интеллекта для производства новых предметных знаний.

В статье [5] было дано определение ***предметно-ориентированных систем научной осведомленности*** как узкоспециализированных систем научной осведомленности, которые кроме возможности решения задач интеллектуального анализа данных наделены способностью производства новых профессиональных знаний.

Целью настоящей работы является описание программно-технологической архитектуры предметно-ориентированной системы научной осведомленности по физической химии радикальных реакций и ее отдельных программных компонентов.

Предметно-ориентированная система научной осведомленности по физической химии радикальных реакций

Разрабатываемая предметно-ориентированная система научной осведомленности по физической химии радикальных реакций рассматривается как интеллектуальная система в интернете, назначением которой является сбор, хранение, верификация, извлечение, распространение и производство новых предметно-ориентированных знаний по физической химии радикальных реакций.

Предметную область системы составляют следующие объекты и их основные характеристики:

- органическая молекула (химическая формула, наименование, характеристика реакционного центра, CAS Number, тип разрываемой связи, энергия диссоциации связи, энтальпия образования молекулы);
- радикал (химическая формула, наименование, характеристика реакционного центра, CAS Number, энтальпия образования радикала);
- радикальная реакция (реагенты, тип реакции, реакционная способность, фаза, растворитель, условия протекания реакции);
- библиографическая ссылка на источник фактографических данных;
- электронная лекция для дистанционного обучения и задачи для самостоятельного контроля;
- тезаурус предметно-ориентированных знаний (определение терминов, контекстно-зависимые подсказки).

Для разработки и создания системы был использован подход на основе портальных технологий, технологии интеллектуальных агентов, технологий прикладного искусственного интеллекта, технологии баз и хранилищ данных. Активным компонентом системы является интеллектуальный агент, который можно представить в виде веб-приложения, расположенного за внешним информационным порталом. При этом сами агенты ориентированы на обработку научных данных в узкоспециализированном разделе предметной области.

Система состоит из нескольких программных слоев. Первый слой реализован как предметно-ориентированное *веб*-приложение, которое предоставляет пользователю интерфейс и принимает управляющие решения на основе таблиц решений. Данное веб–приложение предоставляет доступ к следующим программным компонентам системы: информационной подсистеме, аналитической подсистеме, подсистеме дистанционного обучения, встроенной подсистеме объяснений и подсистеме производства новых профессиональных знаний.

Второй слой состоит из интеллектуальных агентов, реактивных агентов и обученных искусственных нейронных сетей, которые реализуют работу встроенных в портал экспертных систем и выполняют функции поиска информации. Агенты распределены в различных узлах локальной сети.

Третий слой представляет собой хранилище знаний, которое состоит из:

- хранилищ экспериментальных данных по константам скорости радикальных жидкофазных реакций, энергиям диссоциации связей молекул, энтальпиям образования молекул, реализованных как связанные киоски данных (интенсиональные киоски данных);
- производных хранилищ данных по константам скорости радикальных реакций в жидкой и газовой фазах, энергиям диссоциации связей молекул, энтальпиям образования молекул, в которые пользователи имеют возможность заносить полученные ими в результате работы с системой новые значения (экстенсиональные киоски данных);
- базы знаний, которая состоит из правил продукций, таблиц решений, процедур выполнения расчетов, алгоритмов кластерного анализа данных и общих фактов, используемых экспертными системами;

- электронных документов, которые в частности представляют собой материалы учебных лекций по физической химии радикальных реакций, а также тезаурусы терминов и файлы подсистемы объяснений.

Система предназначена для решения следующих задач:

- поиск экспериментальных данных о реакционной способности реагентов в бимолекулярных радикальных реакциях в жидкой фазе;
- поиск расчетных данных о реакционной способности реагентов в бимолекулярных радикальных реакциях в жидкой и газовой фазах;
- поиск значений энергий диссоциации связей органических молекул, а также энтальпий образования радикалов и молекул;
- поиск в базе данных библиографических ссылок;
- оценка реакционной способности реагентов радикальных бимолекулярных реакциях в жидкой и газовой фазах по термохимическим и кинетическим данным;
- оценка энергий диссоциации связей органических молекул по кинетическим данным.

Хранилище знаний

Хранилище знаний по определению есть предметно-ориентированная, интегрированная, поддерживающая временные ряды данных электронная коллекция, которая содержит данные, знания, процедуры генерирования знаний. Оно используется для анализа и исследования данных, производства новых знаний и поддержки принятия решений.

Хранилище знаний настоящей системы содержит эмпирические и рассчитанные факты, продукционные правила и процедуры расчета, а в совокупности с экспертными системами образует виртуальную подсистему производства новых профессиональных знаний (констант скорости и энергий активации радикальных реакций, энергий диссоциации связей молекул).

Хранилище знаний как компонент производства новых профессиональных знаний включает в себя:

- хранилище данных для исследования (Exploration Data Warehouse), содержащее экспериментальные данные по реакционной способности радикальных реакций в жидкой фазе;
- встроенную экспертную систему для управления оценкой реакционной способности реагентов радикальных реакций, представляющих собой совокупность интеллектуального и реактивных агентов;
- интеллектуальный агент для оценки константы скорости и энергии активации реакции в жидкой и газовой фазе;
- веб-сервис, посредством которого осуществляется вызов обученных искусственных нейронных сетей для прогнозирования значений констант скорости и энергии активации жидкофазных радикальных реакций определенных классов;
- производное хранилище данных, содержащее рассчитанные данные по реакционной способности радикальных реакций в жидкой и газовой фазах;
- встроенную экспертную систему для оценки энергии диссоциации связи молекул по кинетическим данным радикальных реакций;
- хранилище данных энергий диссоциации связей органических молекул, которое может быть пополнено новыми данными в результате работы экспертной системы оценки энергии диссоциации связей молекул по кинетическим данным радикальных реакций;
- тезаурус основных терминов и понятий предметной области;

- тезаурус описаний алгоритмов и процедур прогнозирования физико-химических характеристик молекул;
- базу знаний, содержащую продукционные правила и факты, которые используются встроенными экспертными системами.

В результате работы пользователей системы хранилище знаний пополняется новыми профессиональными знаниями. В качестве механизмов производства новых знаний в настоящей системе используются встроенные в портал экспертные системы, обученные искусственные нейронные сети и интеллектуальные агенты. Как показала опытная эксплуатация системы, использование встроенных экспертных систем оправдано в тех случаях, когда возможно использовать знания экспертов или получить такие знания в процессе интеллектуального анализа данных, что является трудоемким по времени процессом. Более перспективным оказалось использование искусственных нейронных систем, которые в некоторых случаях лучше аппроксимируют зависимости в данных и дают более точный прогноз реакционной способности. Однако искусственные нейронные сети не всегда можно обучить из-за отсутствия представительной обучающей выборки.

Структуры данных для хранилища и баз данных

Информационный компонент системы является частью хранилища знаний и состоит из баз и хранилища данных. Хранилище данных системы реализовано как взаимосвязанные киоски данных, построенных методом многомерного моделирования. Информационный компонент системы включает в себя:

- киоск данных констант скоростей радикальных жидкофазных реакций (КД КСРЖФР);
- киоск данных энергий диссоциации связей органических молекул (КД ЭДСОМ);
- киоск данных энтальпий образования органических молекул (КД ЭООМ);
- базу данных по энтальпиям образования радикалов;
- базу данных библиографических ссылок.

Для проектирования структур данных был использован метод многомерного моделирования [6].

Таблица фактов *Rate_constants* содержит фактографические данные о константе скорости радикальной реакции и перечень ключей таблиц измерений. Факты не являются аддитивными и полуаддитивными по всем измерениям. Однако по комбинации измерений к некоторым фактам могут быть применены статистические функции. Будем называть такие факты ***полуаддитивными по статистической функции***. Таким полуаддитивным фактом является логарифм предэкспонента, к которому можно применить статистическое среднее для определенной комбинации измерений, фиксирующей набор реакций с одинаковыми реакционными центрами.

Факты не являются аддитивными и полуаддитивными по всем измерениям. Однако факт - значение энергии диссоциации связи - можно считать полуаддитивным по некоторой комбинации измерений, если необходимо определить среднюю прочность связей в молекуле. Последняя величина используется в некоторых химико-физических расчетах в органическом синтезе.

Логическая структура данных КД ЭООМ аналогична многомерной модели КД ЭДСОМ с той лишь разницей, что те же самые таблицы измерений находятся в логической связи с таблицей фактов, содержащей значения энтальпии образования молекул.

Логическая структура базы данных по энтальпиям образования радикалов включает в себя таблицы измерений *Rad_Compounds, Reference, Method* и таблицу *dH_Value*, содержащую значения энтальпии образования радикала.

Логическая структура базы данных библиографических ссылок представляет собой объединение в виртуальной таблице таблиц измерений *Reference* всех киосков данных.

Производство новых данных

Производство новых данных в рамках предметной области рассматриваемой системы научной осведомленности выполняется с помощью экспертных систем. Экспертные системы (ЭС), предназначенные для функционирования в среде Интернет, конструируются в виде набора интеллектуальных программных агентов – автономных программ с определенным поведением. Под агентом понимается вычислительная система, помещенная во внешнюю среду, способная взаимодействовать с ней, совершая автономные рациональные действия для достижения определенных целей [7].

Абстрактно агент может быть представлен функцией:

$$action: S \rightarrow A, \quad (1)$$

где внешняя среда описывается множеством состояний среды S, а возможные действия агента описываются множеством действий A.

Резидентные агенты являются интеллектуальными агентами. Они обладают своей базой знаний и механизмом вывода для принятия решений. Резидентный агент является, как правило, агентом с состоянием: он обладает внутренней структурой данных, которая может быть модифицирована в зависимости от восприятия текущего состояния внешний среды. Таким образом, текущее состояние внешней среды влияет на выбор действий агента.

Пусть I – множество внутренних состояний агента и P – множество возможных восприятий окружающей среды. Тогда резидентного агента можно представить парой функций: функцией, отвечающей за изменение внутреннего состояния,

$$refine: I \times P \rightarrow I, \quad (2)$$

и функцией действия

$$action: I \rightarrow A. \quad (3)$$

Реактивные агенты - вычислители не обладают своей базой знаний и функционируют по схеме «условие-действие». Они принимают входные данные, обрабатывают их и возвращают ответ резидентному агенту. Действие этих агентов определяется текущим состоянием и может быть представлено функцией (1).

Реактивные обучаемые агенты имеют свою базу знаний и наделены возможностью обучения и расширения своей базы знаний. Обучение агента выполняется в автономном режиме работы с участием эксперта или без эксперта. Возможность использования накопленного опыта агентом может быть представлена функцией:

$$action: S \times A \rightarrow A. \quad (4)$$

Отметим, что условие автономности обучения предполагает пассивность агента в обучении, и такого агента нельзя считать самообучающимся агентом, поскольку он только периодически повышает свою квалификацию.

Поисковый агент предназначен для поиска и извлечения необходимых данных из объектов хранилища знаний и может быть представлен функцией (1).

Программные агенты функционируют в рамках простой модели «Запрос-Ответ-Соглашение». После получения входных данных производится опрос резидентных агентов. На основе полученных ответов принимается решение, какому агенту поручить выполнение предусмотренных в экспертной системе действий. После опроса агентов формируется матрица ответов, на основе анализа которой принимается решение о том, какому агенту отдать решение задачи. При некоторых условиях решение задачи может быть отдано двум агентам.

Экспертная система оценки энергии диссоциации связи по кинетическим данным была реализована на основе рассмотренной выше многоагентной архитектуры и встроена в портал системы научной осведомленности по физической химии радикальных реакций.

Энергия диссоциации связи является одной из фундаментальных характеристик молекулы и влияет на скорость протекания химической реакции, которая описывается совокупностью характеристик: константой скорости реакции, энергией активации реакции, показателем частоты соударений реагирующих ингредиентов и температурой (кинетическими данными).

Экспертную систему обслуживают следующие агенты:

- Агент типа A1 выполняет поиск в базе расчетных данных. Этот агент предлагает свои услуги, если в расчетных данных хранилища знаний по энергиям диссоциации связей имеются данные.
- Агент типа A2 выполняет поиск в экспериментальных данных хранилища знаний. Этот агент предлагает свои услуги, если в экспериментальных данных по энергиям диссоциации связей имеются данные.
- Агент типа A3 выполняет оценку энергии диссоциации связи молекулы на основе эмпирической модели радикальных реакций. Этот агент предлагает свои услуги, если вектор входных параметров содержит достаточно данных для проведения расчета. Оценка реакционной способности может быть выполнена как в жидкой фазе, так и газовой фазе. Алгоритм выполнения оценки изложен в [4].
- Агент типа A4 использует для выполнения оценки энергии диссоциации связи молекулы обученную искусственную нейронную сеть. Этот агент принимает решение об оказании своих услуг, если его обученная искусственная нейронная сеть отвечает заданным входным данным. Агент имеет возможность переобучать свою сеть в автономном режиме.
- Интеллектуальный агент - резидент - выполняет анализ входных данных, выбирает агентов исполнителей, анализирует полученный результат и возвращает его на интерфейс экспертной системы.
- Пользователь с помощью интерфейса экспертной системы может сохранить полученный результат в базе расчетных данных, заполнив специальную анкету проведения расчета.

Для представления знаний в базе знаний интеллектуального агента-резидента используется продукционная модель, то есть знания представляются в виде продукций:

$$(i): Q; P; A \Rightarrow B; N, \qquad (5)$$

где i - имя продукции, Q - сфера применения продукции, P − условие применимости ядра продукции, $A \Rightarrow B$ − ядро продукции, N - постусловия продукции. В базе знаний

продукция представляется в виде таблицы правил и таблицы фактов. Таблица правил содержит ядра продукций в виде пары объектов <условие>-<вывод>. Например, **ЕСЛИ** радикал = алкильный **И** молекул = парафин **ТО** класс = R_1 + R_1H. Таблицы фактов содержат описание параметров класса, параметров радикала и параметров молекулы.

Для представления внутренних состояний и возможных действий в зависимости от текущего состояния внешней среды используется внутренняя структура в виде таблицы решений «состояние–действие». История изменений внутреннего состояния резидентного агента сохраняется в его базе знаний для обеспечения возможности возврата к предыдущему состоянию, если текущие действия агента признаются человеком – экспертом неадекватными.

Экспертная система для управления оценкой реакционной способности реагентов радикальных реакций также разработана на основе многоагентной архитектуры, которая включает в себя резидентного агента, агента вычислителя и набора искусственных нейронных сетей для предсказания реакционной способности реагентов для определенных классов реакций. Предварительно обученные нейронные сети реализованы как отдельные веб-сервисы в распределенной вычислительной среде.

Реакционная способность реагентов в жидкой или газовой фазах определяется, как правило, либо парой значений «температура - константа скорости (скорость взаимодействия реагентов)», либо тройкой значений «интервал температур – энергия активации реакции – предэкспонент».

Интерфейс системы

В главном меню предметно-ориентированной системы научной осведомленности по физической химии радикальных реакций каждый пункт (URL-ссылка) соответствует варианту использования системы. Пункт меню «E&k» отсылает пользователя к подсистеме оценки реакционной способности элементарной реакции, которая включает в себя встроенную экспертную систему и набор искусственных нейронных сетей для производства новых значений констант скорости и энергий активации радикальной химической реакции.

Пункт меню «E&kDB» предоставляет пользователю интерфейс для поиска данных о реакционной способности молекул в радикальных реакций в жидкой и газовой фазах. Киоск экспериментальных данных по константам скорости молекул в радикальных реакций в жидкой фазе насчитывает более 31000 записей. Данные по газовой фазе включают в себя только расчетные значения констант и энергий активаций. База данных по константам скорости молекул в газовой фазе собирается NIST и опубликована на сайте http://kinetics.nist.gov/kinetics/index.jsp.

Пункт меню «$\Delta_fH(R^{\cdot})$» предоставляет пользователям доступ к базе данных по энтальпиям образования радикалов. База данных в настоящее время содержит более 1000 значений. Пункт меню «Δ_fH» предоставляет пользователю доступ к базе данных по энтальпиям образования молекул. База данных в настоящее время содержит более 1000 значений. Содержание обеих баз данных основывается на результатах полученных авторами [7]. Аналогичные базы данных представлены на сайтах NIST (http://webbook.nist.gov/chemistry/) и МГУ им. М.В. Ломоносова (http://www.chem.msu.ru/rus/handbook/ivtan/welcome.html и http://www.chem.msu.ru/cgi-bin/tkv.pl?show=welcome.html/welcome.html).

Пункт меню «BDEDB» предоставляет пользователям доступ к киоску данных по энергиям диссоциации связей органических молекул, который в настоящее время содержит более 1000 записей.

Пункт меню «BDE» отсылает пользователя к экспертной системе для оценки энергий диссоциации связей органических молекул по кинетическим данным радикальных реакций отрыва.

Пункт меню «Bibliography» предоставляет пользователю доступ к базе данных библиографических ссылок.

Пункт меню «Expert» предоставляет доступ к интерфейсу эксперта, который может редактировать данные в базах данных и хранилище знаний.

Пункт меню «e-learning» отправляет пользователя в подсистему дистанционного обучения, которая включает в себя электронные курсы лекций, электронные задачники и тесты контроля знаний в удаленном режиме.

Полученные в результате использования пользователем экспертных систем константа скорости реакции или энергия диссоциации связи органической молекулы могут быть сохранены в хранилище знаний. Такая возможность делает систему активной и позволяет заинтересованному научному сообществу накапливать в ней новые знания.

Возможность пополнения системы предметными знаниями накладывает на ее функционал определенные ограничения. Такие ограничения связаны с необходимостью обеспечить достоверность заносимых в нее данных. Поэтому хранилище знаний системы состоит из двух разделов: базового раздела, составленного экспертами по данным научных публикаций, и раздела, произведенного пользователями системы.

При попытке занесения новых данных экспертная система контекстного управления на основе нечетких рассуждений делает ряд проверок и выводов о достоверности этих данных, а затем принимает решение либо о занесении данных с определенным показателем их надежности либо об отказе в запоминании данных. При сохранении данных система просит пользователя заполнить анкету.

Заключение

Разработка и публикация в интернете предметно-ориентированных систем научной осведомленности на основе использования хранилищ знаний с применением многоагентой технологии позволит научному сообществу создавать распределенные сети для сбора, хранения, извлечения, интеллектуального анализа, распространения и производства новых знаний в узкоспециализированных областях исследований и технологий.

Отметим, что представленная в данной работе система представляет собой самостоятельно функционирующий объект в сети Интернет: она предназначена развиваться через пополнение ее хранилища знаний пользователями и экспертами.

Включение в такие системы подсистемы дистанционного обучения предметно-ориентированным знаниям значительно расширяет круг ее потенциальных пользователей (студентов и аспирантов), что способствует самостоятельному формированию у них профессиональных знаний, а преподавательскому составу высших учебных заведений предоставляет дополнительный учебный материал и электронный ресурс-справочник.

Литература:

1. Hackathorn R. Science Intelligence. Can a Business Intelligence Approach Enable "Smart" Science? DM Review. - 2005. [Электронный ресурс]. – режим доступа: http://www.DMReview.com.

2. Thierauf R.J. Effective Business Intelligence Systems. Westport. Quorum Books. - 2001. - 392 p.

3. Dong Q., Yan X., Chirico R.D., Wilhoit R.C., Frenkel M. Database Infrastructure to Support Knowledge Management in Physicochemical Data // 18-th CODATA Conference. Montreal, Canada, 2002, Sep 29 – 3 Oct. - 36 p.

4. Tumanov V.E. Data Warehousing and Data Mining in Thermochemistry of Free Radical Reactions // Fourth Winter Symposium on Chemometrics "Modern Methods of Data Analysis". Russia. Chernogolovka. February. 15-18. 2005. - P.28-29.

5. Туманов В.Е. Предметно-ориентированные системы научной осведомленности // Информационные технологии. - 2009. - № 5. - С.12-18.
6. Туманов В.Е. Проектирование хранилищ данных для приложений систем деловой осведомленности (Business Intelligence Systems): Учебное пособие / В.Е. Туманов — М.: Интернет-Университет Информационных Технологий: БИНОМ. Лаборатория знаний, 2010. — 615 с.: ил., табл. — (Основы информационных технологий).
7. Wooldridge M., Jennings N. Intelligent Agents: Theory and Practice // Knowledge Engineering Review. - 1995. - № 10 (2). - P.115–152.
8. Денисов Е.Т., Туманов В.Е. Оценка энергий диссоциации связей по кинетическим данным радикальных жидкофазных реакций // Успехи химии. - 2005. - Т. 74.- № 9. - С. 905-938.