Перспективы развития информационно-аналитических средств в задачах сбора и генерации справочных данных.

В.Ю. Зицерман, Г.А. Кобзев, Л.Р. Фокин

1. Введение. Вне зависимости от конкретной области науки, процессы подготовки и распространения справочных данных по свойствам веществ и материалов (СДСВ) включают примерно одни и те же этапы: (1) сбор и компиляция исходных данных; (2) их экспертный анализ с выявлением согласованности, оценкой погрешностей, подбором моделей и т. п. (3) статистическая обработка, связанная с определением значимых параметров модели и оценкой погрешностей рекомендуемых данных; (4) распространение итоговых (рекомендуемых) данных в виде печатных и электронных публикаций и/или заполнение баз данных (БД); (5) использование данных в вычислительных программах, обеспечивающих моделирование природных или технологических процессов.

При ограниченных масштабах работ по подготовке СДСВ потребности в тотальной информатизации рабочего процесса не возникает – использование компьютерных методов охватывает лишь стадии 3 (статистическая обработка данных) и 5 (моделирование процессов). Однако, с нарастанием масштабов работ как по числу веществ и свойств, так и по объемам накопленных и публикуемых данных, охвату библиографии и т. д., справочно-аналитическая деятельность все более приобретает черты производственного, а, правильнее сказать, бизнес процесса, требующего не только надежной научной базы, но и отработанной технологии контроля и управления информационными потоками. В этих условиях потребность в информационных технологиях возникает на всех стадиях справочной деятельности, включая накопление и экспертизу данных, их распространение, координацию совместной работы и проч. Помимо вычислительных средств для решения собственно научных задач, в работе по подготовке СДСВ возникает необходимость в специальных компьютерных методах, способных формализовать и автоматизировать документооборот, обработку текстовой и численной информации, анализ данных, принятие решений и прочие задачи, аналогичные возникающим при управлении корпоративными структурами, вне зависимости от их назначения: производство, финансы, информационный и библиографический сервис и др.

Надо заметить, что в последние годы развитие информационных технологий происходило под влиянием не столько традиционных задач, порождаемых потребностями точных наук (физика, механика, и т.п.), сколько задач, характерных для сферы бизнеса и делопроизводства. Основу всех этих задач составляет работа с обширными массивами данных (численных, текстовых, графических), включая организацию их хранения и обработки. При этом все чаще используются появившиеся в последние годы средства *интеллектуального анализа* данных, дополняющие или заменяющие средства традиционной статистики и нацеленные на поиск закономерностей, аномалий и т. п., то есть, выявление *знаний*, скрытых в *сырых* (необработанных) данных [1-4].

Ориентация информационных технологий на решение проблем из сферы бизнеса стимулировала активное внедрение алгоритмов решения *плохо формализуемых* задач [1], в противовес традиционным методам хорошо алгоритмируемых задач, с которых начиналось развитие вычислительной математики, нацеленной на решение задач физики, механики, теории управления и т.п. Накопленный в последние годы опыт показывает, что *плохо формализуемые задачи* лежат в основе не только бизнеса, но и множества научных дисциплин, изучающих природу и человека: медицины, психологии, общественных наук и т.д. Основные особенности этих задач в том, что неизвестны аналитические зависимости или цепочки действий, приводящие к результату без вмешательства человека, а исходные данные отличаются неполнотой, противоречивостью и искажениями. При этом, на одно из первых мест выходит организация хранения и выборки данных. Только на основе накопленных больших массивов численной (а также текстовой или графической) информации удается отстроить процедуры анализа и обработки, позволяющие выявить и использовать «скрытые» в исходной информации знания.

Для таких задач появились специальные математические средства, реализованные в виде программных кодов: нейронные сети, нечеткая логика, эволюционные (в частности, *генетические*) алгоритмы. Есть специальные программные технологии, объединенные под названием *DATA MINING* [1, 2], которые, используя эту математику, вне зависимости от предметной области, решают задачу интеллектуальной обработки данных с поиском закономерностей, аномалий и прогнозов.

Основная цель данной статьи – рассмотрев особенности и структуру СДСВ, а также рабочий процесс по их подготовке и распространению, обратить внимание на то, что эта деятельность, по сути, нуждается в тех же видах информационных технологий,

которые уже нашли достаточно широкое применение в сфере бизнеса: интеллектуальный анализ данных (с опорой на алгоритмы решения плохо формализуемых задач), электронный документооборот, а также некоторые новые подходы к хранению и распространению данных. Мы попытаемся сопоставить достигнутый в деятельности по СДСВ уровень использования информационных технологий с теми возможностями, которые уже сегодня открывает рынок программного обеспечения.

При этом, авторы опираются на хорошо знакомую им сферу справочных данных по физико-химическим и теплофизическим свойствам веществ и материалов. Нельзя исключить, что, скажем, практика работы с ядерно-физическими константами или данными, используемыми в био- или геоинформатике, может предъявлять несколько иные требования к выбору технологий и программных средств.

2. Логико-информационная структура СДСВ. Общепринятые требования к качеству СДСВ (в отношении достоверности, согласованности, полноты и т. п.), а также условия их использования в БД, сетях и вычислительных приложениях, определяют целую совокупность характеристик, отражающих структуру этого типа данных.

1. Фонд СДСВ включает как первичные (*сырые*) данные, полученные непосредственно из эксперимента, так и вторичные, прошедшие качественную экспертизу и статистическую обработку. При этом полученные в результате обработки вторичные (рекомендованные) данные по объему и структуре могут заметно отличаться от первичных экспериментальных данных.
2. Для целей хранения и распространения применяются многообразные формы представления данных. Первичные данные задаются, как правило, в табличной, реже, в графической форме; вторичные – в табличной форме, в виде аналитических функций (например, уравнений состояния) или программных кодов.
3. Как первичные, так и вторичные данные по свойствам веществ имеют некоторую логическую структуру, строго соответствующую физической модели для веществ данного класса (в простейшем случае структура может

быть представлена в виде таблицы). Однако, для широкого класса справочных данных наблюдаются вариации логической структуры применительно к отдельным наборам данных, например, применительно к разным веществам или разным свойствам. Такая особенность данных позволяет отнести их к, так называемым, *слабоструктурированным* (или *полуструктурированным*) данным, в соответствии с типологией, принятой в теории БД [5]. Изменчивость логической структуры данных означает, что типовая форма в виде перечня *атрибутов* (свойств), приписываемых объекту (веществу, материалу) может заметно меняться в зависимости от класса вещества, его изученности, диапазона параметров и т. д. Например, обычный набор физико-химических свойств, включающий данные по точкам кипения, плавления и критической точке становится «непредставителен» при переходе к высоким температурам, когда в смеси появляются компоненты, не выделяемые в виде отдельного вещества: радикалы, ионы, фрагменты молекул и проч. Другой пример - вещества со сложным составом пара такие, как *S, Se, HF*, N_2O_4; их полное описание требует данных как по веществу в целом, так и по компонентам смеси. При подготовке справочника [6] авторы не раз сталкивались с необходимостью менять форму представления свойств реальных газов в зависимости от их особенностей. Однако, если в печатном издании это требует лишь соответствующих комментариев, то для формализованной структуры БД, интерфейсов и приложений это создает немалые проблемы. Специалистам по БД [1, 5] эта проблема знакома на примере временных рядов, когда приходится отслеживать изменения не только показателей, но и классификаторов, номенклатуры и т. п., с тем, чтобы обеспечить совместное использование данных с изменяющейся во времени структурой.

4. Данные по свойствам в скрытом виде содержат «знания» в виде математической модели, определяющей зависимость свойств от параметров состояния, фазы, вещества (или класса веществ), и «в идеале» оценки достоверности модели. Выделенные в явном виде, эти знания позволяют: (1) провести верификацию данных на предмет их достоверности и согласованности; (2) выполнить экстраполяцию и интерполяцию по параметрам; (3) расширить номенклатуру свойств, доступных для пользователя; (4) построить (или выбрать из предопределенного множества)

математическую модель, то есть, функциональную зависимость свойства (или комплекса свойств) от параметров состояния. Особое значение в технологиях подготовки СДСВ имеет верификация согласованности данных. Так называемая, *совместная обработка данных* (*multiproperty or simultaneous analysis*), когда физическая модель восстанавливается по данным различной природы, что дает одновременный контроль достоверности всей совокупности данных и наиболее надежную оценку параметров модели, стала в настоящее время одной из основных технологий при подготовке СДСВ [7, 8]. Например, потенциал межмолекулярного взаимодействия (базовая информация для молекулярно-кинетической теории) надежно определяют на основе множества разнородных данных: термодинамических, транспортных, спектральных, данных по рассеянию молекулярных пучков и проч. Знание потенциала (его формы и параметров) позволяет создать систему внутренне согласованных данных по второму вириальному и кинетическим коэффициентам для разреженных газов, а в отдельных случаев и для плотного флюида (например, для инертных газов). Анализ согласованности данных включает критический анализ экспериментальных методов, полноты учета погрешностей и т. п.

5. Заключенные в данных физические знания существенно меняют организацию БД или вычислительной системы, используемой как потребителями, так и экспертами, отвечающими за наполнение фондов. Возможность эффективной свертки исходной информации позволяет сократить объем и унифицировать структуру вторичных данных, возложив решение пользовательских задач на приложения. Так вся совокупность рекомендованных термодинамических данных может быть сведена к набору параметров, определяющих уравнение состояния и теплоемкость идеального газа.
6. Активное использование знаний при подготовке СДСВ предполагает работу эксперта с целым множеством моделей, заранее выбранных на основе определенных физических представлений. Например, для обработки теплофизических данных привлекают потенциалы различной структуры в зависимости от типа молекулы (формы, мультипольных моментов, атомного состава) и доступности данных. Точно также при построении PVT-поверхности варьируют ее форму, меняя либо число параметров в заранее

выбранном аналитическом выражении, либо его структуру, добавляя, к примеру, *скэйлинговый* фрагмент для прецизионного описания критической области. В практике обработки данных эти процедуры являются достаточно рутинными и привычными, но именно они предъявляют серьезные требования к построению информационной технологии. Множественность моделей (*структурная оптимизация*), во-первых, усиливает значимость данных, структура которых может меняться от одного экземпляра или набора данных к другому (*плохо структурированные* данные – см. выше). С другой стороны, в задачу БД входит теперь манипуляция не только данными, но и знаниями, то есть подбор моделей, адекватных поставленной задаче: типу вещества, свойства, диапазону параметров и проч. От БД, как основного элемента информационной технологии, происходит постепенный переход к *базе знаний*, способной кодировать накопленные знания о предметной области и продуцировать новые знания [9].

7. Сущность физических знаний, заключенных в первичных данных, означает также, что при выборе определенного свойства для определенного вещества эти данные должны быть согласованы как с известными закономерностями (например, со строгими термодинамическими соотношениями), так и с другими данными, например, по другим свойствам того же вещества, или по тем же свойствам, но других веществ, что должно, например, соответствовать теории термодинамического подобия (теории соответственных состояний).

8. Справочные данные имеют ценность для пользователя только при наличии дополняющей их информации о достоверности, представленной либо в виде оценок погрешностей, либо в более полной форме, включающей *вариационно-ковариационную* матрицу параметрической модели. Информация о достоверности данных является важнейшим фрагментом СДСВ. Без этой информации невозможно: (1) сопоставлять различные наборы данных для сравнительной оценки их качества и надежности; (2) проверять согласованность разнородных данных, скажем, данных по сжимаемости и энтальпии; (3) использовать данные для расчета и моделирования физических процессов; (4) выделять из данных физически значимую информацию, например о критических постоянных или параметрах потенциала. В статье [10] специально подчеркивалось, что

оценка неопределенности выделяет справочные данные из всех разновидностей научно-технических информационных ресурсов, придавая им особое качество: наличие дополнительной информации («знание о незнании»), которая хранится в сознании человека или памяти компьютера и используется в процессе принятия решений. В целом, проблеме оценки достоверности СДСВ посвящена обширная литература, включая нормативные документы Госстандарта [11, 12].

9. Объем, структуру и форму представления справочных данных стараются согласовывать с целями их последующего использования для математического моделирования природных или технологических процессов. Использование СДСВ в задачах моделирования предъявляет к ним целый набор требований в части полноты фонда, формы представления, диапазона параметров, логической структуры данных и проч. Прежде всего, разработчик данных согласует перечень веществ и свойств с физической моделью процесса. Например, для типовых задач технологии основного органического синтеза многолетняя практика позволила выбрать примерно 400 органических и простейших неорганических веществ и около 20 характерных параметров, позволяющих оценивать физико-химические свойства отдельных веществ и композиций [13]. Однако, номенклатура веществ и свойств резко меняется при переходе к задачам высокотемпературной химии: процессы горения, плазмохимия, металлургические процессы и проч. [14, 15]. Своя номенклатура есть в БД для задач металлургии, геохимии, теплоэнергетики. Естественно, что заметно меняется и диапазон параметров в зависимости от предметной области. Например, для химической технологии верхняя температура, как правило, составляет 1000 ^{0}C, в то время, как в термохимических БД для высокотемпературных процессов верхний предел достигает $(20 \div 60) \times 10^3$ К.
10. Весь мировой опыт подготовки СДСВ показывает, что эти данные должны рассматриваться как *динамические*, то есть, изменяющиеся во времени. Массивы СДСВ (как исходные, так и рекомендованные) постоянно обновляются и расширяются в связи с новыми экспериментами, моделями, методами обработки и т. п. Отсюда также следует необходимость перехода к специальным технологиям, апробированным на системах корпоративного документооборота, которые способны обеспечить автоматическую

актуализацию электронных публикаций по мере обновления исходных данных.

3. Перспективы использования новых информационных технологий в работах по подготовке СДСВ. На сегодняшний день арсенал компьютерных средств, используемых при подготовке и распространении справочных данных, в основном, включает: (1) типовые БД для хранения первичных (исходных) и вторичных (рекомендованных) данных; (2) интегрированные среды, объединяющие базы рекомендованных данных и приложения, обеспечивающие расчет физического процесса; наиболее распространенный пример - интеграция программ расчета химического или фазового равновесия с БД по энергиям Гиббса индивидуальных веществ; (3) многообразные вычислительные средства для решения задач статистической обработки первичных данных.

Между тем, проведенный выше анализ структуры СДСВ как специфического информационного ресурса показывает, что оправдано внедрение новых информационных технологий, способных удовлетворить более широкому набору требований, возникающих при подготовке и распространении справочных данных.

Требование *интеллектуализации* средств обработки. Под *интеллектуализацией* средств обработки данных понимается наделение их способностью к выявлению «скрытых» знаний и манипулированию не только данными, но и знаниями (моделями). На сегодняшний день в справочно-информационной деятельности использование развитых технологий управления знаниями (*knowledge manage*) является скорее исключением, чем правилом.

Вообще говоря, соотношение данных и знаний по их объему и значимости сильно меняется в зависимости от предметной области. БД, используемые в химической промышленности, по преимуществу основаны на предсказании всей совокупности макроскопических свойств (термодинамических, переносных, эксплуатационных) по ограниченному объему данных для индивидуального вещества (данных о точке плавления, критических постоянных и т.п.). При этом объем исходных данных невелик, и вся тяжесть справочно-информационной работы переносится на методики прогнозирования свойств: их выбор из большого множества, проверка на адекватность, оценка точности и проч. Встречаются и крайние случаи, когда модель допускает априорное (почти без использования макроскопических опытных данных)

предсказание свойств, как, например, квантово-статистическая модель плотно сжатого вещества [16]. Напротив, в материаловедении типична ситуация, когда нет априорных моделей или корреляций, и опытные данные являются единственно значимым источником информации. Примерно такова же ситуация в таких областях как полимерная химия, фармакология и пр., где в отсутствие работоспособных математических моделей накоплен огромный фактический материал по структуре, свойствам и эксплуатационным характеристикам веществ и материалов. Именно в таких ситуациях оправдано использовать методы интеллектуального или *глубинного* анализа (*DATA MINING*) [1, 2], чтобы из «сырых» данных вычленить закономерности типа «структура-свойство». Интеллектуализация информационных технологий, понимаемая, как их способность к формализованному представлению на равных основаниях данных и знаний, становится, по-видимому, важной тенденцией в работах, связанных с оценкой свойств веществ и материалов. Далее на конкретных примерах мы рассмотрим примеры внедрения интеллектуальных методов в справочно-информационной деятельности.

Требования, связанные с адекватной оценкой неопределенности. Здесь мы выделим лишь те аспекты, которые существенны при выборе информационных технологий. Прежде всего, сведения о достоверности должны включаться как в первичные, так и во вторичные (рекомендованные) наборы данных. Достоверность первичных данных является базовой характеристикой при их обработке, а достоверность вторичных данных позволяет оценить неопределенность технологического расчета или проекта, где они используются. В большинстве Центров по подготовке данных стараются включать в фонды как авторские оценки погрешностей первичных данных, так и те, что получены экспертом при их анализе и/или обработке; экспертиза данных и их проверка на согласованность часто вскрывают неучтенные авторами случайные и систематические ошибки. Сведения о достоверности данных включают погрешности значений, а также качественную информацию: о надежности экспериментального метода, согласии с данными других авторов, степени изученности вещества или класса веществ и проч. Все эти сведения эксперт использует, принимая решения о назначении статистических весов, выборе/исключении отдельных точек или наборов данных и т. п. Последующая процедура обработки, включающая назначенные статистические веса, частично показывает, насколько обоснованы были априорные предположения эксперта. Одновременное использование в работе *нечетких знаний* эксперта (сведения о качестве данных, их надежности и т. п.) и количественных характеристик делает актуальной

задачу перехода к технологиям, способным формализовать работу с качественной информацией. В практике построения информационно-аналитических систем уже давно применяются алгоритмы нечеткой логики (*fuzzy logic*) [17-19], допускающей произвольные значения истинности информации в интервале от 1 (истинная) до (0) (ложная), что позволяет на равных основаниях включать в обработку количественные и качественные оценки достоверности. Есть интересные примеры информационных технологий, использующих при обработке численных данных, в частности, временных рядов, экспертной информации, включающей качественные оценки и суждения [20].

Требования, связанные с использованием СДСВ. Требования, предъявляемые к технологиям работы с данными, как и к самим СДСВ, в значительной степени продиктованы практикой их использования: предметной областью, кругом задач и т. п. Для моделирования предпочтительным является формат представления в виде стандартизованных аналитических функций или программных кодов. Возможны весьма специфические требования к применяемым алгоритмам и *software*, например, к выбору независимых переменных. Так в теплотехнике используют одну из пар термодинамических переменных (температура-энтропия, энтальпия-энтропия и др.*)*, а не температуру и давление, как в большинстве научно-технических приложений. Весьма разнообразны варианты выдачи данных по составу смесей и растворов: мольные доли, числа молей, моляльности и т. п., в зависимости от удобства постановки последующих расчетов. Для многовариантных расчетов иногда приходится отказываться от прецизионных данных в пользу упрощенных моделей, с целью сократить временные затраты. Табличную (или графическую) форму сохраняют лишь для качественного анализа, приближенных оценок, иллюстраций, преподавания и т. п., хотя в задачах материаловедения графическая форма представления диаграмм состояния, по-прежнему, является основной. Есть большая литература [14, 15 и ссылки там], где подробно освещается адаптация БД к вычислительным потребностям в различных предметных областях. Выполнен ряд исследований, позволивших методом Монте-Карло переносить погрешности справочных данных на результат моделирования (например, на неопределенность в оценке КПД, выхода целевого продукта и т. п.) и на этой основе оценить желаемый уровень точности СДСВ [21-23].

Требования, связанные с динамическим характером СДСВ. Как уже отмечалось, динамический характер справочных данных связан с реальной практикой их перманентного уточнения и пересмотра, вызванного расширением массива исходных данных, апробацией новых моделей (например, уравнений состояния), методик

обработки и согласования и т. п. Отсюда следует целесообразность применения в работах по СДСВ, по крайней мере, двух новых (но уже опробованных в сфере бизнеса) информационных технологий: (1) электронные публикации в виде динамических документов (используются также термины *виртуальный* или *динамически компонуемый* документ [24]); (2) электронный документооборот, система, обеспечивающая автоматизацию и протоколирование всех рабочих стадий, повторяемых экспертами по мере появления новых экспериментальных данных или привлечения новых математических моделей, методов обработки и проч.

Сам по себе переход к электронным документам (размещаемым на CD или на сервере), как основному средству публикации данных, имеет ряд преимуществ, в сравнении с традиционной формой печатного издания: простота издательской технологии, включение в документ графических и вычислительных модулей, интеграция с БД или электронными таблицами, отсутствие ограничений на объем текстового и табличного материала. Поставка электронного документа как источника справочной информации открывает ряд уникальных возможностей (особенно для теплофизики), например:

- можно отказаться от работы с обширными таблицами, заменив их аналитическими выражениями или программными кодами, реализующими алгоритмы расчета свойств, (термодинамических, транспортных и т.п.);
- исключаются работы по распространению в текстах числовой информации с большим числом значащих цифр, что неизбежно сопровождается ошибками или невоспроизводимостью результатов на разных вычислительных платформах;
- появляется возможность непрерывного обеспечения пользователей новыми версиями уравнений состояния или корреляций для фазовых равновесий, появляющихся при включении в обработку новых данных;
- появляется возможность представления системы справочных данных в виде динамических сетевых документов, автоматически обновляемых при расширении или ревизии фонда первичных данных.

Переход от обычных (статических) документов, сохраняющихся в том виде, как они были созданы, к динамическим электронным документам - принципиально новый

шаг во всей технологии подготовки и распространения справочных данных. Динамический документ определяют как совокупность объектов (таких как текст, графика, таблица, программный модуль и др.), которая динамически компонуется в единый документ по запросу пользователя в процессе его интерактивной работы с документом [24]. Основная особенность такого документа - распределенный характер информации и ее динамическая сборка в момент запроса со стороны пользователя, хотя пользователь при работе с таким документом непосредственно не наблюдает особенностей его структуры и компоновки. Технология работы с документом, подлежащему динамической сборке при обращении к нему, порождает целый ряд преимуществ, как для пользователя, так и для эксперта, отвечающего за его изготовление:

- для пользователя доступна более «свежая» информация, поскольку при каждом обращении заново генерируется содержание документа;
- для создателя документа легче обновлять содержание, так как обновление любого фрагмента приводит к обновлению всех документов, содержащих указатель на этот фрагмент;
- информация, составляющая динамический электронный документ, является распределенной, то есть может храниться на разных (в том числе, удаленных) серверах;
- реализуется главный принцип современной технологии программирования - *однократное хранение при многократном использовании,* что экономит дорогостоящие ресурсы памяти.

В итоге можно обеспечить перманентную генерацию новых версий таблиц (или уравнений) по мере появления новых опытных данных.

Публикация новых данных ставит и другую задачу – повторное проведение экспертизы и обработки всей совокупности данных (как ранее опубликованных, так и новых) с проверкой их согласованности и достоверности. Типичной является ситуация, когда вновь опубликованные данные требуют от эксперта пересмотреть отношение к старым: изменить оценки погрешностей, исключить отдельные точки или наборы данных, отметить наличие несогласованностей или систематических ошибок. С появлением новых данных может возникнуть задача перехода при статистической

обработке к другой математической модели, например, введению дополнительных слагаемых в уравнение состояния. Абстрагируясь от научной стороны вопроса, можно сказать, что такие процедуры всегда связаны с необходимостью одновременной работы с большим числом документов, в которых фиксируются как сами данные, так и детали, связанные с их обработкой и принятием решений (по оценке погрешностей, выбору модели и т.д.)

Сама процедура ревизии и обработки данных достаточно трудоемка и включает много звеньев, так что желательно для последующей работы фиксировать не только исходные данные и оценки их погрешностей, но и весь протокол рабочего процесса, приведшего к новым оценкам. Если не зафиксировать весь объем этой промежуточной информации, эксперту придется повторять громоздкую процедуру анализа и обработки без учета уже найденных ранее решений. Надо учитывать также необходимость согласования работ на отдельных звеньях технологической цепочки, требования по координации работы различных экспертов, в том числе, удаленных друг от друга и проч. Справиться со всем комплексом подобных проблем, тяжесть которых нарастает по мере увеличения масштабов справочно-аналитической деятельности (числа веществ, количества обработанных публикаций, частоты ревизии данных), может помочь нашедшая в последние годы применение система *электронного документооборота* [24-26]. Под термином *электронный документооборот* понимается формализованная совокупность процедур сбора, хранения, передачи и обработки документов в их жизненном цикле. Основу систем электронного документооборота, который стал внедряться как система ведения архивно-библиотечных средств, процедур поддержки решений и др., является интеграция различных продуктов и технологий в рамках единой среды с постепенным вытеснением «бумажного документа».

Системы электронного документооборота заменяют традиционное ведение БД, позволяя автоматизировать все стадии работы с документом: его ввод, индексирование и каталогизацию, отслеживание рабочего процесса (*workflow*) движения документа между разными экспертами, процедуры аналитической обработки и, наконец, генерацию новых документов с их распространением по потребителям информации.

При этом система следует алгоритмам, регламентирующим движение документа, автоматически фиксируя его модификации после работы каждого исполнителя; в случае со справочными данными система могла бы фиксировать, каким образом и на каком этапе обработки эксперт сделал заключения о погрешности данных. Фиксировать желательно различные промежуточные решения: исключение

отдельных точек или наборов данных, пересмотр априорных оценок погрешностей, указания о несогласованности данных, о наличии систематической погрешности, о методе обработки и т. п.

Из всего спектра задач по манипуляции документами, решаемых системами электронного документооборота, технологии подготовки СДСВ наиболее близка задача управления потоком работ (*workflow*) [5, 24]. В основном, она включает, так называемую, систему маршрутизации и контроля исполнения. Разработано несколько подходов к маршрутизации рабочего процесса: *документо-ориентированный* (маршрутизируется документ как основной объект, а остальные параметры маршрутизации ассоциируются), *работо-ориентированный* (основным объектом является работа, к которой прикрепляют разнообразный список документов и приложений), жесткая маршрутизация документов по заранее определенным маршрутам с контролем исполнения, свободная маршрутизация с назначением маршрута по результатам движения документов и ряд др.

Поддержание и хранение протоколов движения документов, если они включают наборы данных и параметры моделей, необходимы для оценки результатов работы эксперта, при пересмотрах ключевых физических величин, учете изменений температурной шкалы и многих других задач. Специфика справочной деятельности, как уже отмечалось, состоит в том, что работа по ревизии и обновлению фонда практически не прекращается, в связи с перманентным появлением новых данных, физических моделей, расширением диапазона параметров и т. п. Хранение и использование всей информации о процедурах предыдущего этапа позволяет облегчить и формализовать последующие работы по ревизии и обновлению данных.

Подобные протоколы могут помочь также в решении относительно новой проблемы, с которой столкнулись специалисты по компьютерной обработке данных: принципиальной *невоспроизводимости* результатов при статистической обработке с использованием многопараметрических моделей [27].

Создателем «дисциплины воспроизводимых исследований» является сейсмолог из Стэнфорда Джоун Клейрбаут (http://sepwww.stanford.edu/research/redoc/cip.html), который обнаружил, что даже сотрудники его лаборатории не могли воспроизвести собственные результаты обработки данных, спустя всего лишь полгода после завершения проекта. Практически невозможным оказывается восстановить все сочетание параметров алгоритмов и версий программ. К 1991 году Клейрбаут с сотрудниками разработали программный комплекс, для генерации так называемого

воспроизводимого электронного документа (ReDoc); последний включает в свой состав как сам отчет об исследовании, так и специальные утилиты, обеспечивающие использование нужных версий исходных текстов при создании выполняемого файла под требования дисциплины воспроизводимых исследований. Аналогичные виды электронных документов разработаны также для пакета MATLAB, универсальной среды проведения математических и инженерных расчетов. Протокол потока работ, по сути, может решать ту же задачу, если в дополнение к сведениям по достоверности будет включать информацию о программном обеспечении. Уместно отметить также важнейшую проблему сохранения исходной и промежуточной информации при смене поколений научных работников.

Требования по работе с плохо алгоритмируемыми задачами. Переход от хорошо алгоритмируемых задач, с которых начиналось развитие вычислительной техники (математическая физика, теория управления, задачи оптимизации и проч.), к задачам плохо формализуемым, которые лежат в основе бизнеса и большинства дисциплин, изучающих природу и человека (науки о Земле, медицина, психология, общественные науки) в последние годы становится одной из основных тенденций в развитии информационных и вычислительных технологий. Работа со СДСВ, понимаемая в широком смысле, относится именно к плохо формализуемым задачам, поскольку в этой деятельности приходится использовать большие массивы неполной и искаженной количественной информации, качественные оценки, нечеткие сведения и т.п. Работа со СДСВ является столь сложной и многозвенной именно из-за принципиально плохой формализуемости многих этапов, связанных с оценкой качества данных, их проверкой на полноту, согласованность и достоверность [7]. В основе их анализа лежит обработка массивов данных с многочисленными искажениями: систематическими и случайными ошибками, пропусками, промахами и т. п. [28 - 30]. При анализе необходимо учесть неформализуемые сведения об уровне эксперимента (надежность и полнота данных, их соответствие известным закономерностям и пр.). Результатом обработки являются замкнутые аналитические выражения, составляющие основу достоверного знания о свойствах вещества, что в конечном итоге позволяет резко сократить объем рекомендуемых данных, сохраняя только константы аналитических формул. Хотя статистическая обработка в различных вариантах является основной процедурой при подготовке данных, практически используются традиционные методы нелинейной регрессии. При этом экспертные оценки влияния неполноты информации, несовершенства эксперимента и т.п. не совсем строго включаются в оценки

статистических весов и не совсем последовательно переносятся на оценки погрешностей результатов. Здесь имеется открытое поле деятельности для специалистов по математической статистике.

4. Современный уровень компьютерных технологий в работе Центров подготовки данных. Перечислив те требования к информационным технологиям, которые связаны со спецификой справочно-аналитической деятельности, интересно на ряде характерных примеров выяснить, насколько удовлетворяются эти требования в крупных Центрах по подготовке данных, для которых внедрение новых технологий наиболее актуально как из-за масштабов их деятельности, так и за счет достаточно заметных финансовых ресурсов, которые они могут выделить на цели информатизации. Оценить степень использования новых технологий в работе над справочными данными можно по многочисленным публикациям, содержащим описание БД и/или интегрированных сред, обеспечивающих обработку и распространение данных по свойствам, а также из материалов WEB-серверов, поддерживаемых Центрами данных. Относительно подробный обзор дан в публикациях авторов [31] и в недавней монографии [15].

Хотя конкретные технологии различаются достаточно сильно, необходимые заключения можно сделать по описанию наиболее распространенных систем, например БД Термодинамического исследовательского Центра (TRC) США [32] или крупных систем, интегрирующих несколько БД и вычислительных средств для расчета фазовых равновесий и построения диаграмм состояния, таких как F*A*C*T (Facility for the Analysis of Chemical Thermodynamics) и ThermoCalc.

Система БД TRC, одного из наиболее известных в мире Центров по физико-химическим свойствам органических веществ, может служить ярким примером того, как при высоком уровне внедрения компьютерных технологий, их возможности далеко не в полной мере используются, чтобы обеспечить требуемую эффективность процедур обработки и распространения численных данных. В полном соответствии со спецификой СДСВ пользователи TRC получают доступ к двум основным БД: исходных экспериментальных данных (БД SOURCE) и обработанных рекомендованных данных (TRC TABLE). Однако, вся процедура превращения «сырых» данных в рекомендованные скрыта от пользователя, и каких либо сведений об ее протоколировании и последующем использовании не приводится. Распространяются

данные TRC по сети ИНТЕРНЕТ в виде рабочих файлов для БД ACCESS, расположенной на компьютере пользователя. Это значительно более эффективный способ распространения в сравнении с традиционным (текст, включающий таблицы), поскольку открывает множество возможностей поиска данных, их экспорта в табличной или графической форме, формульной аппроксимации и т.п. Файлы MS ACCESS не являются *динамическими документами* (как и обычный справочник) в том смысле, что отражают структуру и объем данных на момент создания, и с появлением новых источников не подвергаются перестройке; пользователь в этом случае должен запрашивать на сервере новую версию БД. Эта же особенность присуща и всем БД, распространяемым на CD, вне зависимости от конкретных форматов данных и файлов.

Надо сказать, что создатели БД TRC отчетливо видят необходимость перехода к *динамически компонуемым* документам, позволяющим перманентно и без участия пользователя обновлять фонд с появлением новых исходных данных и повторением процедур обработки. Авторы [32] предложили создать распределенную систему хранения и пополнения архива теплофизических данных, позволяющую любому из пользователей сгенерировать в момент запроса динамический документ, отражающий тот объем данных и знаний, который доступен на момент обращения к системе. В основу проекта положена современная технология БД с удаленным доступом, средствами проверки данных, копирования и восстановления, способностью работать на разных платформах. Предполагается, что таблицы экспериментальных данных должны быть «слинкованы» с полнотекстовыми документами, содержащими исходную публикацию.

Большое место в проекте уделено, так называемым, *метаданным* («данные о данных») [1, 5]. Метаданные позволяют проводить автоматический отбор и обработку данных, различая, например, прямые результаты эксперимента или сглаженные данные из опубликованных отчетов. Для дистанционной загрузки опытных данных разработана специальная программа LOADER2, которая проверяет внутреннюю согласованность исходной информации и дополняет отсутствующие данные приближенными оценками. Напомним (см. раздел 2, п. 3), что СДСВ с точки зрения логико-информационной структуры относятся к *слабоструктурированным* данным [5], что требует при организации их хранения в БД специальных технологий, основу которых составляет именно использование метаданных, которые разъясняют смысл и формат числовой информации.

Сочетание программы LOADER2 и дистанционно пополняемой БД TRC SOURCE составляет основу технологии генерации динамических документов для представления СДСВ. Заметим, однако, что эта технология пока не получила развитие и намечена лишь как перспективная.

Достаточно слабо представлена линия на интеллектуализацию БД с постепенным ее превращением в базу знаний. Каждое из рекомендованных свойств представлено в отрыве от других, без проверки разнородных термодинамических данных на внутреннюю согласованность. Использование интегрирующего начала в виде единого уравнения состояния, функции Гельмгольца или потенциала межмолекулярного взаимодействия, могло бы быть первым шагом к переходу от разрозненных данных к системе знаний, описывающих закономерности поведения вещества и прогнозирующих широкий спектр числовых характеристик. Другой возможный шаг в переходе от данных к знаниям – сочетание экспериментальных данных с прогнозными методиками, например основанными на групповых вкладах, учитывающих особенности структуры молекулы, также (согласно [32]) намечен лишь как необходимая тенденция развития.

В работах Центра не поставлена задача документирования и управления процессам подготовки новых данных, по-видимому, с учетом того, что используются довольно тривиальные процедуры, без построения многопараметрических моделей, согласования и отбраковки данных и т. п. Переупрощены и структуры данных, практически не выходящие за пределы типовой табличной формы, что не предполагает активное использование *метаданных* для изменчивых логических структур, которые постоянно встречаются при охвате широкого круга соединений или привлечении множества физических моделей. При использовании достаточно производительных БД, таких как ORACLE и многоярусной системы активных сетевых серверов, уровень обработки и представления данных не сильно изменился в сравнении с 70-80 годами прошлого века, и в целом заметно ниже, чем в работах с другими видами численных данных(например, с каталогами астрофизических и географических данных, данными биоинформатики и т. п.), не говоря уже о многочисленных сферах использования деловой и производственной информации [1, 5, 9, 24]. Заметим при этом, что речь идет об одном из наиболее авторитетных в мире Центров данных, обеспеченном техникой, программным обеспечением и кадрами специалистов по информационным технологиям.

Анализ упомянутых выше систем (F*A*C*T, ThermoCalc), решающих задачи вычислительной термодинамики (расчет равновесий, построение диаграмм состояния) демонстрирует те же ограничения в использовании логико-информационных и интеллектуальных ресурсов информационных технологий: преимущественная работа с рекомендованными данными типовой структуры, крен в сторону использования приложений, а не данных, упрощенные процедуры обработки без документирования рабочего процесса, без формализации экспертного анализа и т. п. Разумеется, речь идет лишь о тенденциях, а в конкретных разработках БД можно найти попытки решения упомянутых проблем. Так ведущий в России Центр данных, ТЕРМОЦЕНТР им. акад. В.П. Глушко, наряду с машинными средствами хранения и распространения справочных данных, имеет сложную человеко-машинную систему, обеспечивающую сбор, хранение и экспертно-статистический анализ первичных данных о молекулярных постоянных, теплотах реакций и термодинамических функциях веществ в стандартном состоянии [6, 15]. В БД по фазовым диаграммам полупроводниковых систем [33] весьма широко представлена исходная информация наряду с полученной в ходе расчета и согласования данных, но сама процедура обработки данных выведена за рамки БД и не доступна пользователю. В системе ЭПИДИФ [34], обеспечивающей хранение и обработку данных по теплофизическим свойствам разреженных газов, пользователю выдается согласованная система данных с оценками погрешностей этих свойств.

Один из немногих примеров технологий, используемых для интеллектуализации работ со СДСВ дает система *Cranium* [35], для которой основной задачей является свободное манипулирование молекулярными структурами и методиками оценки физико-химических свойств, аналогично тому, как БД манипулирует численными и текстовыми данными по свойствам. В отличие от прикладных программ, работающих с фиксированной структурой данных, база знаний подбирает программный код, реализующий методику оценки из обширного фонда, обеспечивая по отношению к методикам типовые операции, такие как добавление, удаление, редактирование, копирование и т. п. Объектно-ориентированная технология позволяет системе *Cranium* на равных условиях управлять как данными (например, данными о молекулярной структуре, точкам кипения и плавления, термохимических константах и проч.), так и алгоритмами (методиками), каждый из которых трактуется в системе как объект, с которым допускаются те же действия, что и с фрагментами данных. Объектно-

ориентированная технология обеспечивает простоту создания и поддержки базы знаний, содержащей сотни методик оценки. Для каждой методики хранится своеобразный набор данных, называемый *преамбулой*. Преамбула содержит код, который проверяет данные по структуре молекулы или по составу смеси, чтобы определить применимость методики к данному объекта. Например, многие из методик не способны предсказывать свойства полярных или ассоциированных соединений или не пригодны для водородсодержащих смесей. В случае применимости, преамбульный код дает сведения о точности методики. Таким образом, база знаний пригодна для манипулирования исходными данными, методиками (моделями), и сведениями о применимости или достоверности той или иной модели

Аналогичная задача по созданию средств и технологий автоматизированного доступа к информационному фонду и к математическим моделям предметной области решалась в работах Сергиевской А.Л. [36, 37] для проблем физико-химической кинетики. Система была предназначена для обеспечения информационных и вычислительных задач газодинамики и кинетики, решаемых Центром данных АВОГАДРО (Ин-т механики МГУ). Созданная система обеспечивала согласованное манипулирование информационным фондом системы АВОГАДРО (БД ЧАСТИЦА и ПРОЦЕСС) и структурированными математическими моделями физико-химических процессов.

Для этих целей было дано формализованное описание предметной области с выделением сущностей ЧАСТИЦА→ПРОЦЕСС→СРЕДА и система представления знаний, основанная на целевом преобразовании структурированных наборов данных при помощи используемых в системе математических моделей. Классификация информационных элементов предметной области выполнена в соответствии с разработанной технологией (так называемое, *инфологическое* описание), выделяющей *атомарные* (неделимые) информационные элементы и *агрегации* или *обобщения*, порождающие *составные* элементы. С учетом сложности логической структуры предметной области, отличающейся большим количеством информационных элементов, разнообразием и сложностью связей между ними, выбрана реляционная модель данных, позволяющая одним и тем же набором средств манипулировать данными и связями.

При создании системы манипулирования знаниями были намечены пути решения задач по использованию экспертного знания в виде отдельных суждений, формулировок моделей процессов, оценок работ экспертов и т. п., которые позволяли в

структурированной форме отображать сложившиеся представления по отдельным вопросам формализации знания. В целом, логико-алгоритмическая вычислительная схема предметной области совместно с концептуальной схемой информационного фонда позволяют анализировать явные знания о предметной области и представлять знания в формализованном виде, допускающем манипулирование, аналогичное манипулированию данными. Реляционный вариант представления физико-химических знаний в виде нормализованных таблиц включен в систему управления БД (СУБД) наряду с базовыми таблицами информационного фонда.

Частично решает задачи интеллектуализации БД описанная в работе [38] динамическая библиотека для корреляции физических и термодинамических свойств. Библиотека состоит из программы оценки свойств и БД, которая содержит оригинальные экспериментальные данные и модели регрессий, а также статистическую и графическую информацию для оценки их качества. Программы оценки свойств и БД могут непрерывно модернизироваться за счет включения современных моделей регрессии, экспериментальных данных и регрессионных методов. На примере корреляции теплоемкости для твердых веществ, где экспериментальные данные включают однородные и переходные области, продемонстрированы преимущества применения структуры библиотеки динамических физических свойств над традиционными библиотеками корреляций. Помимо широкого манипулирования моделями и методиками обработки, предложенная библиотека демонстрирует также эффективность использования динамических структур, обеспечивающих передачу пользователю перманентно обновляемых данных.

Как уже отмечалось, процесс интеллектуализации информационных технологий состоит как в свободном манипулировании знаниями (наряду с данными), так и в использовании специфических средств интеллектуального (глубинного) анализа данных, включающих новые вычислительные методы. Из новых алгоритмов работы с «плохими» (искаженными или зашумленными) данными в работе по свойствам веществ, в основном, нашли применение только нейросетевые алгоритмы [1, 17]. Считается, что нейронные сети позволяют максимально использовать доступную информацию при ограниченном объеме экспериментальных данных. Другое их достоинство – возможность аппроксимации в задачах очень большой размерности, при том, что точность аппроксимации при таких методах не зависит от размерности. Имеется довольно много публикаций, авторы которых применяли нейронные сети для традиционных задач физической химии и теплофизики: обработка данных, оценка

свойств по большому массиву данных, установление корреляций типа «структура-свойство» (см., например, [39-42]).

Другой класс методов, активно применяемых сейчас для крупномасштабных и плохо формализуемых задач в сфере бизнеса и технологии – генетические и вообще эволюционные алгоритмы. Они применяются везде, где необходим перебор вариантов с выбором наилучшей из альтернатив. В задачах обработки данных – это выбор модели, то есть подбор функционального выражения, наилучшим образом воспроизводящего физические зависимости, и определение параметров модели по совокупности разнородных данных, причем во многих случаях речь идет о нахождении с высокой точностью нескольких десятков или сотен параметров. Известно, что традиционные методы нелинейного программирования, использующие представления о поведении функции в окрестности экстремума, были мало эффективны при наличии многих локальных экстремумов, оврагов и других особенностей поверхности, что резко ограничивало возможности решения многоэкстремальных задач. В основе новых (эволюционных или генетических) методов лежат формализованные принципы, имитирующие естественный эволюционный процесс, за счет сочетания элементов случайности и детерминированности, точно так, как происходит в природе. Детерминированность состоит в моделировании природных процессов отбора, размножения и наследования по строго определенным правилам. В качестве случайного элемента используется аналог процесса мутации, когда характеристики решения изменяют случайно, чтобы найти новое направление в процессе эволюции решения. Генетические алгоритмы решают задачи, работая с популяцией из некоторого числа наугад взятых решений, которые по аналогии с дарвиновской «борьбой за существование»: скрещиваются (*crossover*), порождают разнообразных «детей», соперничают за ограниченные ресурсы, мутируют, и в конечном счете, умирают. Появилась обширная литература, описывающая методологию и практику эволюционных вычислений [1, 43].

Разработчики СДСВ достаточно редко обращаются к новым методам вычислений, хотя специфика области, казалось бы, дает для этого все основания. Широко использует эволюционные методы статистической обработки группа Вагнера и Спана [44], одна из авторитетных групп по подготовке уравнений состояния и термодинамических таблиц. Используемый ими алгоритм определяет наиболее подходящую форму математической модели (здесь это уравнение состояния, описывающее PVT-поверхность) путем выбора наилучшей комбинации выражений из

корзины регрессоров, обширного математического набора, включающего все мыслимые слагаемые, которые могли бы быть включены в модель. В процессе работы алгоритм комбинирует детерминистские элементы регрессионного метода с процедурами эволюционных вычислений такими, как мутация и оптимизация популяций. Всего с использованием этого алгоритма было построено до 20 уравнений состояния для компонентов воздуха, инертных газов, углеводородов и др. На этом пути удается на 50% сократить размерность задачи в сравнении с традиционными методами нелинейной регрессии при той же точности и добиться лучшей способности модели к экстраполяции.

5. Заключение. Анализируя состояние работ по СДСВ, интересно сопоставить, насколько практика использования здесь новых технологий соотносится с некоторыми из общих тенденций развития информационного общества. В числе этих тенденций можно назвать охват информационными технологиями всех сторон человеческой деятельности, введение безбумажного документооборота, размещение в сети научно-технических и деловых ресурсов, активное использование электронных публикаций и т. п. Все эти тенденции налицо и в деятельности Центров, занимающихся подготовкой СДСВ. Еще одна интересная тенденция — переход от хорошо алгоритмируемых задач (традиционные задачи физики и техники) к задачам плохо формализуемым, характерным для сферы бизнеса, медицины, психологии и множества других дисциплин, изучающих природу и общество.

Проведенный выше анализ показывает, что работа по подготовке СДСВ, будучи традиционной формой научной деятельности, по характеру обработки информации имеет много общего с бизнес процессами, для которых характерны именно плохо формализуемые задачи, требующие совместного учета цифровых и качественных характеристик, нечетких суждений, а также глубинного анализа массивов искаженных данных на предмет выявления закономерностей. Поэтому актуальной задачей становится включение в инструментарий ученых, занятых подготовкой данных, новых технологий (инженерия знаний, нечеткая логика и пр.), прошедших хорошую апробацию на задачах, поставленных в сферах бизнеса и производства, и значительно выходящих по возможностям за рамки стандартных процедур вычислительной математики и статистики.

До сих пор эти средства слабо востребованы в практической работе, а большинство Центров по подготовке и распространению данных ориентируются на традиционные методы, хотя широко используют БД, пригодные для хранения, как исходного материала, так и рекомендованных данных. По-видимому, здесь играют роль традиции, сложившиеся в естественнонаучных коллективах: строгая формализация задач, использование классических методов статистики, игнорирование расплывчатых и субъективных оценок, составляющих суть экспертного анализа. Новые технологии и присущие им математические средства пришли из мира бизнеса, где до их появления обработка данных с выявлением знаний и закономерностей была просто невозможной. Есть и некоторые особенности, отличающие СДСВ от прочих информационных ресурсов с точки зрения логической структуры, требований к процессу обработки, форме представления и т. п., скажем требования к согласованию разнородных данных, перманентное повторение процедуры обработки и проч., что не позволяет воспользоваться готовым программным обеспечением без его перестройки и адаптации к научным задачам. Потребуются дополнительные усилия, чтобы обеспечить активную миграцию новых технологий в естественнонаучную среду, возможно, и не только для поддержки работ по СДСВ.

Работа выполнена при финансовой поддержке РФФИ, Проект № 02-07-90138.

Литература

1. Корнеев В.В., Гареев А.Ф., Васютин С.В., Райх В.В. Базы данных. Интеллектуальная обработка информации. - М.: Изд-во "НОЛИДЖ", 2001 – 351 с.
2. Дюк В., Самойленко А. DATA MINING: учебный курс. – СПб: Питер, 2001 – 366 с.
3. Забежайло М.И. Интеллектуальный анализ данных - новое направление развития информационных технологий. // НТИ. Серия 2, Информационные процессы и системы. – 1998. - №5. - С. 6.
4. Загоруйко Н.Г. Прикладные методы анализа данных и знаний. – Новосибирск: Изд-во Ин-та математики СО РАН, 1999 - 268 с.

5. Когаловский М.Р. Энциклопедия технологий баз данных: Эволюция технологий. Технологии и стандарты. Инфраструктура. Терминология. – М.: Финансы и статистика, 2002 - 798 с.
6. Гурвич Л.В. и др. Термодинамические свойства индивидуальных веществ. Справочное изд. В 4 томах. Под ред. акад. В.П. Глушко. - М:НАУКА, 1978-1982.
7. Фокин Л.Р. Методика создания справочных данных о теплофизических свойствах веществ и ее реализация на примере свойств рабочих тел и конструкционных материалов в энергетике. Автореф. дисс. на соискание уч. степени д.т.н. – Москва: ОИВТ РАН, 1990 – 33 с.
8. Barker J.A. Interatomic potentials for inert gases from experimental data. In “Rare gas solids”, V. 1 - NY, 1976. – P. 1.
9. Гаврилова Т.А., Хорошевский В.Ф. Базы знаний интеллектуальных систем. – СПб: ПИТЕР, 2001 – 382 с.
10. Фокин Л.Р. Справочные данные в системе научно-технической информации //. НТИ. сер. 1. – 1987. - №5. – С. 4-6.
11. Фокин Л.Р., Козлов А.Д., Рабинович В.А., Карпова Г.А. Методика оценки достоверности справочных данных о свойствах веществ и материалов //. Измерительная техника. – 1988. - №11. – С. 7-9.
12. Разработка и аттестация нормативно-справочных данных о свойствах важнейших веществ и материалов. Обзорная информация. Госстандарт СССР; ВНИЦ МВ - М.: Изд-во стандартов, 1987 - 48 с.
13. Reid R.C., Prausnitz J.M., Poling B.E. The Properties of Gases and Liquids. 4th edn., - NY: McGraw-Hill, 1987- 742 p.
14. Байбуз В.Ф., Зицерман В.Ю., Голубушкин Л.М., Чернов Ю.Г. Химическое равновесие в неидеальных системах. Под ред. В.С. Юнгмана. - М.: ИВТАН, 1985. – 227 с.
15. Белов Г.В. Термодинамическое моделирование: методы, алгоритмы, программы.- М.: НАУЧНЫЙ МИР, 2002.- 181 с.
16. Никифоров А.Ф., Новиков В.Г., Уваров В.Б. Квантово-статистические модели высокотемпературной плазмы, методы расчета росселандовых пробегов и уравнений состояния. - М.: Наука: Физматлит, 2000. -399 с.
17. Круглов В.В., Дли М.И., Голунов Р.Ю. Нечеткая логика и искусственные нейронные сети. - М: Физматлит, 2001 - 224 с.
18. Алтунин А.Е. Модели и алгоритмы принятия решений в нечетких

условиях: [монография] - Тюмень: Изд-во Тюменского гос. ун-та, 2002 - 265 с.

19. Ткаченко Т.Я. Интеллектуально-информационная поддержка нечетких наукоемких технологий. Автореф. диссертация на соискание уч. степени д.т.н. - Екатеринбург, Южно-Уральский гос. Университет. - 2002 - 36 с.
20. Головченко В.Б. Прогнозирование временных рядов по разнородной информации -Новосибирск: НАУКА, 1999 – 88 с.
21. Macchietto S., Maduabeuke G., Szczepanski R. Exact Determination of Process Sensitivity to Physical Properties. // Fluid Phase Equilibria. –1986. V. 29. - №1. – P. 59-67.
22. Vasquez V.R., Whiting W.B.Effect of systematic and random errors in thermodynamic models on chemical process design and simulation: A Monte-Carlo approach. // Industrial and Engineering Chemical Research - 1999. - V. 38. -№8. - P. 3036.
23. Whiting W.B., Vasquez V.R., Meerschaert M.M. Techniques for assessing the effects of uncertainties in thermodynamic models and data.// Fluid Phase Equilibria. - 1999. - V. 158. -№1. - P. 627.
24. Клименко С.В., Крохин И.В., Куц В.М., Лагутин Ю.Л. Электронные документы в корпоративных сетях. - М.: "АНКЕЙ" – "Эко-Трендз", 1999 – 271 с.
25. Ларин М.В. Управление документацией и новые информационные технологии. Всероссийский научно-исследовательский институт документоведения и архивного дела. - М. : ВНИИДАД : - Научная книга, 1998 -136 с.
26. Романов Д.А., Ильина Т.Н., Логинова А.Ю. Правда об электронном документообороте. - М.: ДМК, 2002. - 224 с.
27. Левкович-Маслюк Л., «Воспроизводимое и невоспроизводимое» // КОМПЬЮТЕРРА. - 2002 - №3.
28. Петрищев В.С. Анализ плохих данных: (статистический анализ технико-экономической информации) - Обнинск: ГЦИПК, 1998.- 121 С.
29. Литтл Р.Дж.А., Рубин Д.Б. Статистический анализ данных с пропусками. - М: "Финансы и статистика", 1991 – 333 с.
30. Johnston L.P.M.,Kramer M.A. Estimating state probability distributions from noisy and corrupted data.//AIChE J. 1998 - V. 44 - N3 - P. 591.

31. Трахтенгерц М.С., Зицерман В.Ю. Ресурсы ИНТЕРНЕТ для теплофизиков и теплоэнергетиков. Методическое пособие. - М.:ИВТ РАН. Часть 1. Препринт №8-424, 1998, - 58 с. Часть 2. Препринт №8-444, 2000, - 60 с.
32. Yan X., Dong Q., Frenkel M., Hall K.R. Window-Based Applications of TRC Databases: Structure and Internet Distribution // Int. J. Thermophys. – 2001. – V.22. № 1. – P. 227-241.
33. Христофоров Ю.И., Хорбенко В.В., Киселева Н.Н. и др. База данных по фазовым диаграммам полупроводниковых систем с доступом из ИНТЕРНЕТ // Материалы электронной техники. – 2001. - №1. - С. 50-53.
34. Fokin L., Popov V., Kalashnikov A. et al. Joint Russian and Bulgarian Academies of Sciences Database of Intermolecular Potentials and Diffusion Coefficients for 366Components of the CVD Processes in Microelectronics // Int. J. Thermophys.- 2001.- V. 22. - №5. – P.1497-1506.
35. Joback K.G. Knowledge bases for computerized physical property estimation // Fluid Phase Equilibria. – 2001. - V.185. №1-2. – P. 45-52.
36. Сергиевская А.Л., Ковач Э.А., Лосев С.А. Опыт информационно-математического моделирования в физико-химической кинетике. – М.: Изд-во МГУ , 1995. - 311 с.
37. Сергиевская А.Л. Информационное и математическое моделирование физико-химических процессов в газах. Автореф. дисс. канд. физ-мат.н. -Москва: Ин-т математического моделирования, 1993. – 26 с.
38. Shacham M., Brauner N. A dynamic library for physical and thermodynamic properties correlations. // Industrial and Engineering Chemical Research. -2000. - V. 39. -№6. – P.1649.
39. Normandin A. et al. PVT data analysis using neural network models. // Industrial and Engineering Chemical Research – 1993. – V.32. - №5. –P. 970.
40. Lee M.L., Jui-Tang Chen. Fluid Property Prediction with the aid of neural network. // Industrial and Engineering Chemical Research – 1993. – V.32. - №5. –P. 995.
41. Bunz A.P., Braun B., Janowsky R. Quantitative structure-property relations and neural networks: correlation and prediction of physical properties of pure components and mixtures from molecular structure // Fluid Phase Equilibria. – 1999. - V.158. - №1. – P.367.
42. Гальберштам Н.М. Моделирование свойств и реакционной способности органических соединений с использованием искусственных нейронных сетей.

Автореферат на соискание уч. степени к.х.н. – Москва: МГУ, химический факт. - 2001 – 163 с.

43. Батищев Д. И. Генетические алгоритмы решения экстремальных задач. Под ред. Львовича Я.Е.: Учеб. пособие. - Воронеж, 1995. – 69 с.
44. Span R., Collmann H.-J., Wagner W. Simultaneous Optimization as a Method to Establish Generalied Functional Forms for Empirical Equations of State // Int. J. of Thermophysics. - 1998. - V. 19. - №2. - P. 491-500.