УДК 004.318

А.В. Палагин, С.Л. Кривой, Н.Г. Петренко

Знание-ориентированные информационные системы с обработкой естественно-языковых объектов: основы методологии и архитектурно-структурная организация

Рассмотрена совокупность задач, включающая в себя методологический, онтологический и логический аспекты проектирования знание-ориентированных информационных систем, функционирование которых опирается на автоматизацию процессов извлечения и формализации содержания естественно-языковых объектов с последующей обработкой формализованного представления этого содержания логико-семантическими методами с ориентацией на конкретную предметную область.

A set of tasks is considered including methodological, ontologic and logic aspects of designing the knowledge-oriented information systems the functioning of which leans on the automation of the processes of extraction and formalization of the content of naturally-language objects with the subsequent processing of the formalized representation of this content by logically-semantic methods with orientation to a concrete subject domain .

Розглянуто сукупність задач, яка складається з методологічного, онтологічного та логічного аспектів проектування знання-орієнтованих інформаційних систем, функціонування яких спирається на автоматизацію процесів добування і формалізації сенсу природномовних об'єктів з наступною обробкою формалізованого подання цього сенсу логіко-семантичними методами з орієнтацією на конкретну предметну область.

Введение. Не вызывает сомнения тот факт, что интеллектуальные информационные технологии (ИИТ) и соответствующие компьютерные системы (КС) будут интенсивно развиваться в направлениях, соответствующих наиболее существенным признакам разумной деятельности, в том числе распознавания, преобразования и понимания знаковых систем (включая естественно-языковые). Результатом указанных комплексных задач является генерация совокупности смыслов, заложенных в анализируемых знаковых системах. При этом под смыслами понимается то, что делает знаковые системы текстами. В данном случае под термином «текст» понимается любая осмысленная знаковая система, а такой текст уже является источником знаний.

Развитие теории и создание систем искусственного интеллекта (ИИ) в настоящее время связывают с проблемой организации знаний о мире в виде определенных структур, отражающих реальные связи и отношения между предметами и явлениями в окружающей среде, и успехи в этом направлении во многом определяются интеллектуальным уровнем и общей эффективностью технических систем [1, 2].

Постановка задачи

Особенность создания современных информационных технологий состоит в стремлении к максимальной интеграции результатов двух областей ИИ, которые когда-то развивались параллельно и независимо: *Knowledge-engineering* и компьютерной лингвистики (когнитивной семантики). Можно сказать, что это стремление отражает в общем случае естественную схему взаимодействия человека с окружающим миром: «входная информация – сознание – понимание – знания». В последней не указаны промежуточные (ощущение, рефлексия и др.) процедуры когнитивного цикла, чтобы выделить «конструктивную» триаду. Сознание здесь выполняет роль персонифицированного инструмента, вырабатывающего совокупность предметно, ситуативно или причинно-связанных сущностей, образующих «сознательную» картину мира. Выработка указанных сущностей, построение и использование на их основе картины мира реализуется как результат восприятия и преобразования входной информации в когнитивном цикле, ядром которого является *механизм понимания*. Сознание человека имеет языковый статус, и это выражено понятием

«языковое сознание», а сознательная картина мира при этом представлена Языковой Картиной Мира, которая есть не только системой знаний, а и некоторым буфером, связывающим общие знания с профессиональными. Для понимания естественного языка (ЕЯ) в прагматическом плане следует выделить соответствующий механизм, который характеризуется как «...преобразование представления на естественном языке в логические выражения».

Анализ проблемы обработки знаний

Теория и практика создания и использования систем, основанных на знаниях (или знание-ориентированных информационных систем (ЗОИС)), – наиболее актуальное и интенсивно развивающееся направление *Computer Science*, использование результатов которого повысит эффективность создания инструментальных средств, прикладных систем и применение компьютеров. К сожалению, в настоящее время даже на мировом уровне не получено ощутимых и общепризнанных результатов в широких прикладных областях знаний, хотя можно отметить имеющиеся удачные решения в этом направлении для узкоспециализированных приложений. Сложность указанной проблемы определяется, в частности, сложностью организации и использования больших баз формализованных знаний, а также привлечения целого ряда научных теорий (логики, компьютерной и психологической лингвистики, нейрокибернетики, теории семантических сетей и др.), которые, вполне очевидно, должны способствовать решению проблемы извлечения, формального представления, обработки и системной интеграции знаний и составить концептуально-методологическую основу теории междисциплинарных научных исследований.

Как следует из сказанного, исходной информацией для большинства научных исследований служат знаковые системы, представленные в виде естественно-языковых объектов (ЕЯО). В общем случае такими ЕЯО могут быть большие базы неструктурированных данных, хранящихся в корпоративной памяти, разного рода электронные библиотеки, коллекции документов, текстовая составляющая простого и семантического *Web*, монографии, научные статьи, научно-технические и деловые документы и пр. Такое уточнение сужает круг научных дисциплин и подходов, методы которых привлекаются к анализу и пониманию ЕЯО, извлечению и формализации из последних знаний, их структурированию и обработке. Среди прочих можно выделить *лингвистику*, которая исследует, в том числе лексический и структурно-грамматический аспекты ЕЯ, и *логику*, которая рассматривает язык в одном ограниченном аспекте лишь в той мере, в какой он является средством *фиксации*, *переработки и передачи* знаний. Рассматривая языковые описания, логика в первую очередь интересуется отношением его элементов к обозначаемым объектам и тем, как при помощи определенных связей из этих элементов образуются сложные знаковые системы, выражающие истинные знания об объективном мире. Отправляясь от анализа ЕЯ, логика рассматривает в качестве приоритетных особые языки – искусственные языки науки, возникающие на базе естественных, но отличающиеся рядом важных особенностей. Для исследования языков (естественных или искусственных) логика использует инструментарий, реализующийся в виде формальных знаковых систем, позволяющий обнаружить законы построения и функционирования, т.е. образования и преобразования систем знаний. Выделение объективных средств фиксации знаний и выведение одних «единиц» знаний из других по определённым, объективно значимым правилам, является одним из приоритетных в современных исследованиях по логике [3].

При создании интеллектуальных информационных систем (ИИС) следует выделить *три аспекта исследований* – онтологический, логический и методологический. Эти аспекты имеют свои, в общем случае фиксированные объекты исследований, соответствующие процессу познания или разработке некоторой ИИС. В связи с этим все объекты могут быть поделены на три группы: система сущностей (или объектов реального мира), система зна-

ний и система обработки сущностей в соответствии с данной системой знаний. Первая группа является предметом онтологического исследования, вторая – логического исследования и третья – методологического. При этом под методологией будем понимать совокупность приемов, методов и механизмов их взаимодействия, применяемых в процессе исследований.

Сущностный анализ триады «смысл–знание-языковое сознание»

Рассматривая знания как нечто объективное, феноменальное, отражающее *исходную*, *преобразующую* и *конечную* составляющие процесса познания («знание посредством знаний преобразуется в новое знание»), заметим, что совокупность научных дисциплин, имеющих непосредственное отношение к процессам мышления, понимания, осознания и получения новых знаний, не может обходиться без выработки своих собственных понятий, определенным способом фиксирующих свойства и закономерности ее объектов. Из таких основополагающих понятий, как *смысл*, *знание*, (*знаковая*) *система*, *текст*, *объект*, *отношение*, *предмет*, *язык*, *структура*, *связь*, рассмотрим только первые два, выражающие собой наибольшую проблемность исследований в области построения систем ИИ, в том числе и знание-ориентированных информационных систем, оперирующих с ЕЯО.

Известно, что теории смыслов в завершенном виде не существует, а потому и общепризнанного определения понятия «смысл» также нет, как нет однозначного понимания и с точки зрения его формального представления и преобразования в КС. В [3–6] этот термин вводится с гносеологических позиций, через его свойства и атрибуты, его взаимодействие с другими объектами. В [6] отмечается, что «…природа смысла может быть раскрыта только через одновременный анализ *семантической триады*: «*смысл–текст–язык*». Текстовое раскрытие смысла происходит через те знаковые системы, которые мы готовы воспринимать как языки. Таким образом, каждый элемент упомянутой триады раскрывается через два других. Включая в триаду язык, вносим представление о том, что сама триада становится возможной, только когда есть *Наблюдатель* – носитель сознания, воспринимающий тексты и оценивающий смыслы. Триада становится синонимом сознания», скажем точнее – языкового сознания.

Под термином «***смысл***» заданного высказывания либо текста будем понимать функцию интерпретации конечной последовательности символов, в рамках априори согласованной естественно- (либо формально-) языковой семантики (предполагается наличие *Наблюдателя* как носителя языкового сознания).

Известно достаточно много определений понятия «знания» [4, 7–9], дающих некоторое обобщенное представление о системе, основанной на знаниях, но все же этого недостаточно, чтобы построить указанную систему. Для этого необходимо также, по крайней мере, знать:

- чем знание отличается от подобных ему понятий – данных и информации;
- какими общими свойствами обладают знания;
- какие существуют источники знаний;
- какие существуют способы представления и обработки знаний в КС;
- какие существуют механизмы выявления новых знаний.

Поэтому в задачах информатики можно (и более удобно) определить категорию *знания* косвенно – через его свойства и методы обработки.

Знания – это системно зафиксированная в сознании человека (компьютера) совокупность фактов, существенно отражающих реальность материальных и абстрактных объектов окружающего мира.

Под ***системой*** (компьютерных) ***знаний*** будем понимать некоторую конструктивную среду, представленную в базисе подходящего формально-логического языка и обеспечивающую постановку и решение в ней задач из заданной предметной области.

Формальная постановка задачи анализа ЕЯО

Пусть $T = t_1, t_2, ..., t_n$, естественно-языковый текст в алфавите X, т.е. $T \in L(X)$, где $L(X)$ – язык над алфавитом X, а $t_i \in T$ – предложения, $i = \overline{1, n}$, n – мощность множества T.

Каждое предложение $t_i \in T$, в свою очередь, имеет структуру $t_{i_1}, t_{i_2}, ..., t_{i_m}$, где t_{i_j} содержательно означают грамматические единицы, из которых построено предложение t_i. Если $t_{i_j} \in t_i$, то $C_L(t_{i_j}) = t_{i_1}, t_{i_2}, ..., t_{i_{(j-1)}}$ и $C_R(t_{i_j}) = t_{i_{(j+1)}}, t_{i_{(j+2)}}, ..., t_{i_m}$ будем называть *левым* и *правым контекстом* t_{i_j} соответственно.

С текстом T свяжем такие объекты:

- S – словарь языка $L(X)$, где содержатся слова t_{i_j} со своими определителями (в частности, лингво-семантическими характеристиками единиц словаря);
- $\gamma \subseteq T \times S$ – отношение, определяющее возможные значения и типы слова в словаре S;
- $A = (D, \Pi)$ – предметная модель, на которой интерпретируется текст T;
- $\phi \subseteq T \times A$ – отношение интерпретации текста T на области D.

Сигнатура предикатов $\Pi = \{\pi_1^{k_1}, ..., \pi_r^{k_r}\}$ содержит атомарные предикаты, из которых можно строить сложные формулы. Зафиксируем эту сигнатуру, поскольку она зависит от предметной модели. Так как модель не уточняется, то и сигнатуру уточнить нельзя. Заметим только, что каждый атомарный предикат имеет тип, (т.е. речь идет о некоторой типизированной сигнатуре).

Определим правила вычисления отношений γ и ϕ.

Отношение γ имеет достаточно простой способ вычисления:

$$\gamma(t_{i_j}) = \{(a_1, \tau_1), (a_2, \tau_2), ..., (a_s, \tau_s)\},$$

где a_i – возможные значения слова t_{i_j}, а τ_i – его возможные типы. Возможен случай, когда $\gamma(t_{i_j}) = \varnothing$. Тгда значение этого слова считается неопределенным (и это требует пополнения словаря S).

Отношение ϕ определяется несколько сложнее.

$$\phi(T) = \phi(t_1), ..., \phi(t_n),$$

где

$$\phi(t_i) = \{\phi(\gamma(t_{i_1})\gamma(C_R(t_{i_1}))), \phi(\gamma(C_L(t_{i_2}))\gamma(t_{i_2})\gamma(C_R(t_{i_2}))), ..., \phi(\gamma(C_L(t_{i_n}))\gamma(t_{i_n}))\},$$

при этом

$$\phi(\gamma(t_{i_j})) = \gamma(\phi(t_{i_j}));$$

$$\phi(\gamma(C_L(t_{i_j}))) = C_L(\phi(\gamma(t_{i_j}))));$$

$$\phi(\gamma(\pi_r^k(p_1, ..., p_k))) = \gamma(\phi(\pi_r^k))(\phi(\gamma(p_1), ..., \gamma(p_k))),$$

где $\gamma(\phi(\pi_r^k))$ – имя предиката, тип которого согласован с аргументами $\gamma(p_1), ..., \gamma(p_k)$.

Из этой формальной постановки проблемы анализа ЕЯО вытекает, что основные задачи сводятся к таким:

- конкретизировать предметную модель A; эта задача – основная, поскольку предметная модель есть по существу базой знаний (конкретизация состоит в том, чтобы определиться с формальным логическим языком, правилами вывода, аксиоматикой и пр.);
- показать вычислимость отношений γ и ϕ на предметной модели A;
- построить алгоритмы вычисления отношений γ и ϕ;
- при вычислении отношений γ и ϕ контролировать соответствия типов аргументов и предикатов;
- определить взаимодействие алгоритмов вычисления γ и ϕ с системами лингвистического анализа текста.

Второстепенными, но также важными, являются задачи, связанные

- с определением структуры данных для словарей;
- с определением информации, которая должна содержаться в словарях;

• с определением режима взаимодействия с пользователем (автоматический, полуавтоматический или диалоговый).

Анализ взаимодействия базовых понятийных структур в когнитивном цикле обработки знаний

Приведенные описания терминов «смысл» и «знание», а также анализ позволяют синтезировать некоторую обобщенную схему «эволюционирования» знаний (рис. 1). Заметим при этом, что в языке данной статьи словоформы терминов «знание», «сознание» и «познание» имеют общий корень, а сами термины также образуют триаду, компоненты которой фиксируют объект и субъект и процесс познания.

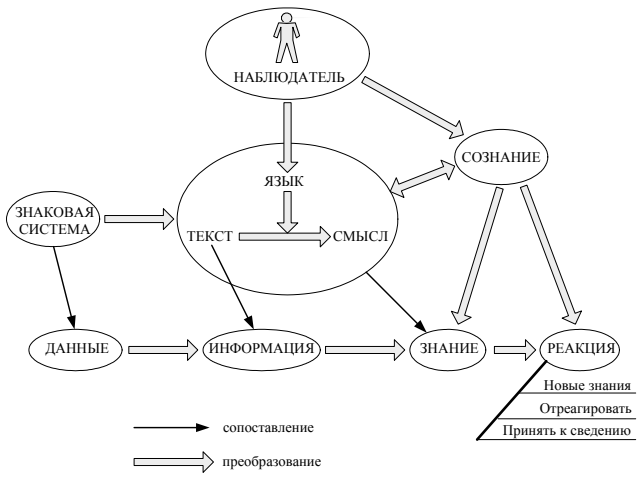

Рис. 1. Обобщенная схема «эволюционирования» знаний

Входом является некоторая знаковая система, которую можно интерпретировать как совокупность данных. Далее происходит «осмысление» входных данных, для чего обязательно присутствие *Наблюдателя* (человек, компьютер) как носителя высших форм мышления – языкового сознания. И если у него имеется «внутренний языковый интерпретатор, настроенный на данную знаковую систему», то данные превращаются в информацию. Далее в работу включается «смысловой интерпретатор» (если он имеется у *Наблюдателя*), который превращает информацию в знания или происходит осознание поступившей информации. На заключительном этапе обработки входного сообщения происходит естественная реакция человека: принять к сведению, отреагировать или пополнить базу знаний (БЗ) новой порцией знаний [1]. В подтверждение сказанного можно привести известный пример Клауса [3]: «Космонавты застают на Марсе надписи – следы вымершей цивилизации. Отмечая некоторую регулярность и пользуясь основными законами математической логики и исчислением вероятностей, они могут получить некоторые знания относительно обнаруженных записей, например, что это математические предложения. Они могут даже реконструировать в некоторой степени знания, заключенные в этих предложениях. Но все это, разумеется, возможно лишь при условии, что они располагают некоторыми универсальными логическими правилами построения и чтения конечных знаковых систем, знакомы с теорией вероятностей и т.п. В противном случае обнаруженные надписи оказались бы совершенно изолированными системами, и говорить о заключенных в них знаниях было бы невозможно».

Свойства знаний

Известно [9], что знания характеризуются рядом свойств, отличающих их от традиционных моделей данных. К таким свойствам можно отнести следующие.

Внутренняя интерпретируемость. При хранении знаний в памяти интеллектуальных информационных систем (ИИС) наряду с традиционными элементами данных хранятся информационные структуры, позволяющие интерпретировать содержимое соответствующих ячеек памяти.

Структурированность. Знания состоят из отдельных информационных единиц, между которыми можно установить классифицирующие отношения: род – вид, класс – элемент, тип – подтип, часть — целое и т.п.

Связность. Между информационными единицами предусматриваются связи различного типа: причина — следствие, одновременно, быть рядом и др. Данные связи определяют семантику и прагматику предметной области.

Семантическая метрика. На множестве информационных единиц, хранимых в памяти,

вводятся некоторые шкалы, позволяющие оценить их семантическую близость. Это позволяет находить в информационной базе знания, близкие к уже найденным.

Актуальность. Данное свойство подчеркивает принципиальное отличие знаний от данных. Выполнение тех или иных действий в ИИС инициируется состоянием базы знаний. При этом предполагается, что появление новых фактов и связей может активизировать систему.

Кроме того, свойство актуальности знаний может порождать процесс актуализации, который имеет определяющее значение в цепочке переходов «хранилище_данных → востребованная_информация → знания». Однако более актуальной (с практической точки зрения) является проблема не первичности компонент цепочки, а их принадлежность. Известны следующие описания принадлежности указанных компонент, признаваемые многими исследователями.

Данные – зафиксированные объекты или явления реального мира.

Информация – знаки, полученные при преобразовании данных в сознании человека или в процессоре компьютера.

Знания принадлежат внутреннему миру человека (компьютера). В полном объеме только в сознании человека происходят сложные и простые когнитивные процессы – от сложных доказательств теорем до силлогизмов Аристотеля.

Для пояснения сказанного можно привести известный пример из школьной практики [8]. Изучение отдельных школьных курсов по соответствующим учебникам (востребованная информация) происходит на определенных этапах, выбираемых из всего множества курсов (хранилища данных), определенных школьной программой. Школьные учителя управляют процессом преобразования информации, предлагаемой школьными учебниками (читаемых школьниками чаще всего «от сих до сих»), в знания. При этом при опросе учеников учителя профессионально при помощи специальных вопросов выясняют, находится ли в голове у школьника запомненная информация (т.е. он урок зазубрил) или она преобразовалась в знание.

Следовательно, на любом этапе преобразования требование актуализации соответствующей компоненты является обязательным (но не достаточным) условием перехода в более высокую степень абстрактности. И если для первого перехода («хранилище_данных → востребованная_информация») механизмы актуализации достаточно хорошо изучены и проработаны, то для второго перехода («востребованная_информация → знания») указанные механизмы являются слабо изученными и реализованы в разрозненных экспериментальных разработках.

Проблема извлечения знаний из текста представляется не только не тривиальной, но и весьма сложной, несмотря на несомненные достижения *Computer science* в этой области. Как справедливо замечено в [8], «Текст есть знаковая конструкция и часто содержит знание. Но текст есть не знание, а только его источник. Знания из текста еще нужно извлечь. Человеку или КС. Библия содержит много знаний, но всякий извлекает их по-своему и не все, что оттуда можно потенциально извлечь».

Источники знаний и проблема формирования новых знаний

В многочисленных исследованиях по тематике ИИ, и в частности работы со знаниями [1, 2, 5, 7, 9–17], выделяют два основных *источника знаний* – это эксперты, специалисты и лингвистический корпус текстов (или множество ЕЯО). Если для первого источника методы приобретения знаний достаточно хорошо изучены и проработаны, а также известны соответствующие промышленные экспертные системы, то для второго – разработаны только отдельные методы, не связанные в единую интегрированную технологию, а соответствующие информационные системы носят экспериментальный характер и не совершенны. Извлечение и обработка знаний из ЕЯО является одним из разделов *Data mining* и признано перспективным междисциплинарным направлением исследований.

Известно достаточно много моделей представления знаний и соответствующих им методов обработки в КС [8–10, 13, 16, 18 и др.]. Их выбор для конкретного приложения существенно зависит от характера знаний ПдО, наличия средств автоматизации построения БЗ, объема последней, а главное – набора реализуемых соответствующей системой функций.

Следует отметить, что существующие механизмы извлечения новых знаний оперируют только знаниями силлогистического типа. Переход к формированию новых знаний более высокого уровня абстракции методами ИИ представляет собой сложную научную проблему. Такие знания, если принять изложенную в [6] вероятностную модель человеческого мышления, соответствуют уровню предсознания *Наблюдателя* и организованы в сложную *онтологическую структуру* с полным набором связей между концептами («Вот где востребована наиболее полная система семантических отношений, связывающая между собой «доступные» *Наблюдателю* понятия реального мира!»). Из сказанного следует, что в настоящее время автоматическое построение БЗ для широких предметных областей представляет сложную научную проблему. (Такое предположение основывается также на известных утверждениях исследователей Винограда, Флореза и Вейзенбаума [13] о том, что наиболее важные аспекты естественного интеллекта в принципе невозможно смоделировать с помощью формальных символьных манипуляций. Необходимы, в частности, обучение и понимание ЕЯ). При этом извне (по отношению к КС) указываются уровни расположения базовых понятий ПдО и базовые отношения между ними. Такую структуру, созданную вручную инженером по знаниям на основе тезауруса ПдО, можно представить как *начальную онтологию ПдО*. При этом не следует забывать об ее лингвистическом аналоге, так как переход от терминов (выраженных в тексте лексемами) к понятиям ПдО далеко не тривиален.

Как отмечено в [9], наиболее перспективными представляются ЗОИС, функционирование которых опирается на автоматизацию процесса извлечения и формализации содержания естественно-языковых текстов с последующей обработкой формализованного представления этого содержания логико-семантическими методами с ориентацией на конкретную ПдО.

Цель данной работы – разработка методологических основ проектирования и архитектурно-структурной организации ЗОИС с обработкой ЕЯО (такая система, разрабатываемая с учетом онтологического подхода, названа *онтолого-управляемой информационной системой* (ОУИС)).

Методологические основы проектирования ОУИС

В соответствии с целью работы главными составляющими методологии являются онтолого-информационно-логический (или инфологический) подход (сочетающий в себе логико-информационную и онтологическую компоненты) к проектированию класса онтолого-управляемых информационных систем, парадигма двойственного участия концептуальных (онтологических) знаний при языковой и проблемной обработке информации, а также виртуальная парадигма (в частности, архитектура компьютерной системы, ориентированная на технологию реконфигурируемого процессинга). Последнюю связывают также с реализацией принципа адаптивности, предполагающего наличие в ОУИС потенциальных возможностей улучшения работы в условиях априорной и текущей неопределенностей на основе обучения и опыта.

Общая схема компьютерной обработки знаний

Общая схема компьютерной обработки знаний, содержащихся в ЕЯО, представлена на рис. 2. Здесь отражена информационная модель обработки знаний, начиная с поиска на всем информационном пространстве текстовых данных заданной в запросе пользователя (возможно в виде предложения на ЕЯ) текстовой информации и последующего ее преобразования (в общем случае) в простые, ситуационные и новые (формальные) знания.

Под *простыми знаниями* понимается поступившая новая информация о некоторой сущности (сущностях) реального мира, что соответствует реакции человека – «*принять к сведению*».

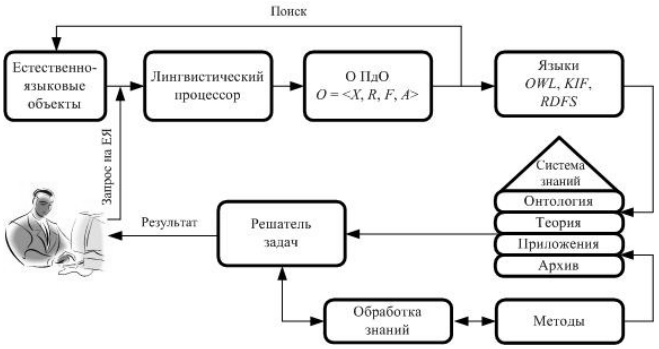

Рис. 2. Компьютерная обработка знаний, содержащихся в ЕЯО

Под *ситуационными знаниями* понимается описание некоторой ситуации, которую необходимо распознать и соответствующим образом «отреагировать».

Под *новыми знаниями* понимается распознавание приращения ΔS (из формулы Брукса $(K(S) + \Delta I = K(S + \Delta S)$, где $K(S)$ – исходные знания, ΔI – новая порция информации, $K(S + \Delta S)$ – выходные знания) и пополнение ими БЗ заданной ПдО в некотором формализованном виде.

Схема классификации средств и методов обработки ЕЯО

На рис. 3 представлена онтолого-классификационная схема концептуальных понятий, верхние уровни которой в совокупности отображают проблематику (структуру) многочисленных научных исследований в области ИИ [5, 14 и др.], а нижние уровни представляют собой объект, предмет и методы исследований при решении научно-технической проблемы построения ОУИС [1, 9, 19–21 и др.].

Здесь приняты следующие сокращения: Т – технология, Иск – искусственный, С – система, Инт – интеллект, Инф – информация, ИТ – информационная технология, ИИ – искусственный интеллект, ИС – информационная система, ПдО – предметная область, ИИС – интеллектуальная информационная система, КЛ –

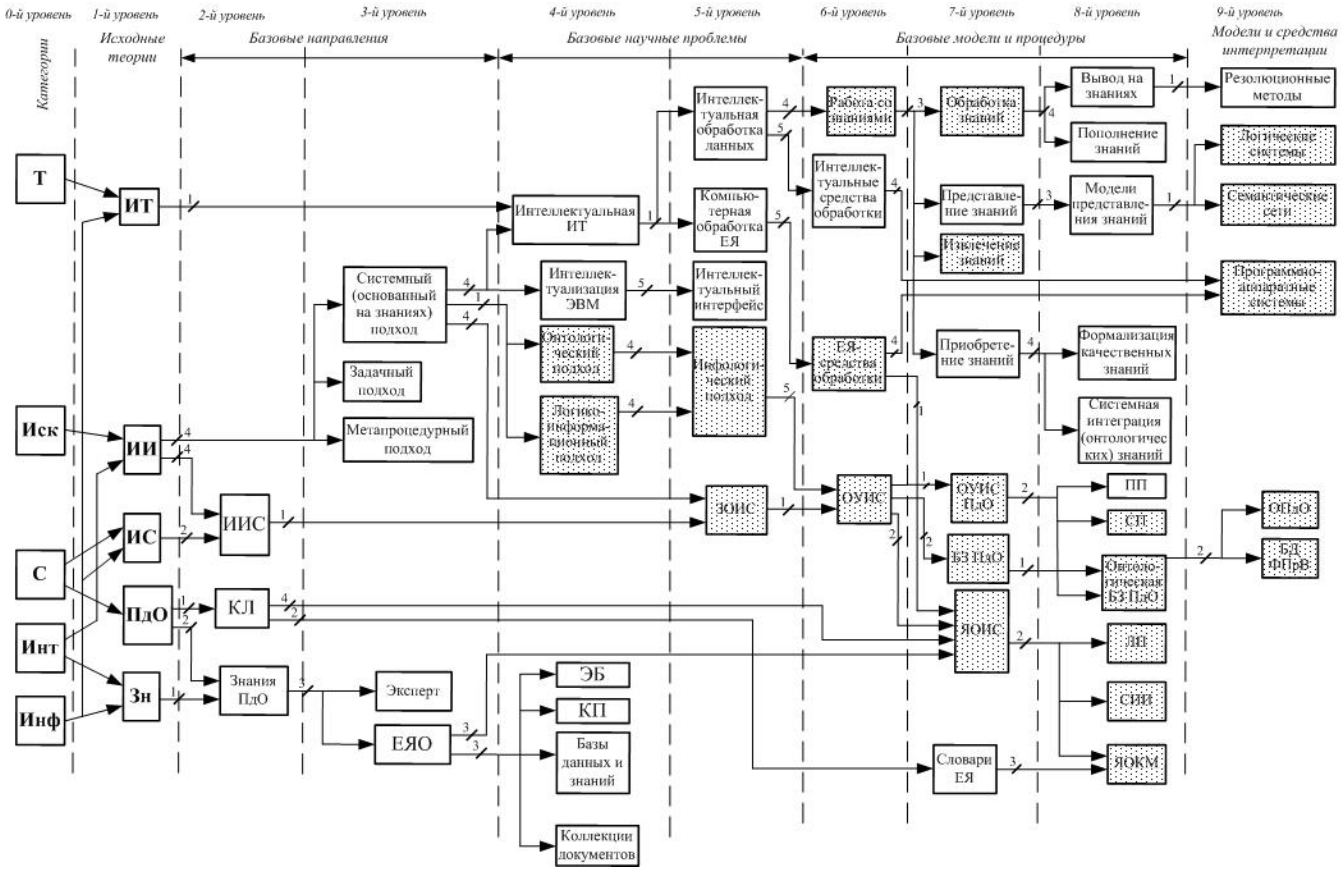

Рис. 3. Онтолого-классификационная схема средств и методов ИИ

компьютерная лингвистика, ЕЯО – естественно-языковые объекты, ЗОИС – знание-ориентированная информационная система, ЭБ – электронные библиотеки, КП – корпоративная память, БЗ ПдО – база знаний предметной области, ОУИС ПдО – онтолого-управляемая информационная система предметной области, ЯОИС – языково-онтологическая информационная система, ЕЯ – естественный язык, ОПдО – онтология предметной области, БД ФПрВ – база данных фактов и правил вывода, ЯОКМ – языково-онтологическая картина мира; типы родо-видовых отношений: 1 – род–вид, 2 – часть-целое, 3 – по объекту преобразований, 4 – по процессу преобразований, 5 – по результату преобразований.

На рис. 4 представлен детализированный фрагмент схемы (рис. 3) с признаками деления понятий (и соответствующей группировкой) по объекту и предмету исследований, в котором отражена совокупность научно-технических задач, входящих в круг решения указанной проблемы. Приняты следующие обозначения: СИИ – семантико-информационный интерпретатор, ЛП – лингвистический процессор, ГА – грамматический анализатор, СА – семантический анализатор, ПП – прикладной процессинг, ОБЗ ПдО – онтологическая БЗ ПдО, ИП – инфологический подход, ФЛП ПЗ – формально-логические методы представления знаний, ФКЗ – формализация качественных знаний, МПЗ – модели представления знаний, АЛС – алгебро-логические системы, СС – семантические сети, КГ – концептуальные графы, *KIF – Knowledge Interchenge Format*, ЛППП – логика предикатов первого порядка. МИ – методы интерпретации, РМ – революционные методы.

На рисунке показаны только концепты и связи между ними, существенные при разработке методологии проектирования ОУИС.

«Онтологизированный» подход к проектированию ОУИС

Проектирование ОУИС предполагает разработку двух взаимодействующих подсистем, соответственно для обработки знаний в заданной ПдО и обработки текстов на основе «языковых» знаний (или знаний из ПдО «Компьютерная лингвистика»). Указанные информационные системы на рис. 3 и 4 представлены соответственно как онтолого-управляемая ИС обработки знаний в заданной ПдО (ОУИС ПдО) и языково-онтологическая ИС обработки текстовой информации на основе языковых знаний (ЯОИС). В общем случае ЯОИС выполняет «языковую» обработку текстовой информации, а ОУИС ПдО – «машинную» обработку формализованных предметных знаний. Взаимодействие между указанными подсистемами осуществляется при реализации некоторого полуформального отображения, использующего базу языковых знаний (в основе которой лежит языково-онтологическая картина мира (ЯОКМ)) [1, 21] и базу знаний заданной ПдО.

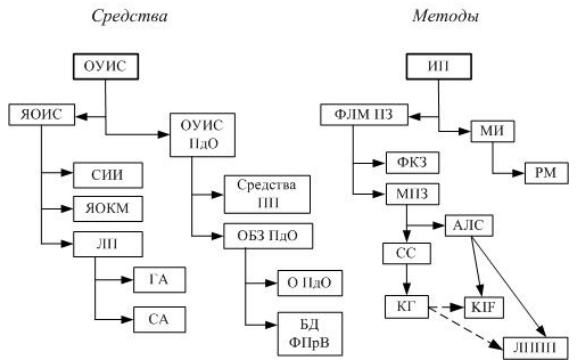

Рис. 4. Средства и методы обработки ЕЯО

Известно, что в процессе создания ЗОИС главными задачами выступают разработка методов и подходов систематизации знаний, что предполагает разработку формальной теории, в том числе языка представления знаний (ЯПЗ), средств обработки (интерпретации) знаний и механизмов вывода новых знаний.

Неотъемлемая компонента любой ЗОИС – наличие в ней базы знаний, где в компьютеризированной форме представлены знания ПдО. Именно наличие этих знаний в системе позволяет ей успешно решать как традиционные, так и новые задачи, которые ранее были исключительно прерогативой человека. В процессе проектирования БЗ все чаще применяется онтологический подход, когда ядро БЗ ПдО есть онтология предметной области (ОПдО). Возник-

ло даже самостоятельное ответвление в исследованиях по построению баз знаний – «*онтологические базы знаний*». Такая онтология (или онтологические знания) одновременно может выступать и как информационная структура концептуальных знаний ПдО, и как один из главных компонентов ЗОИС. Указанное функциональное сочетание отражает два аспекта проектирования ЗОИС – информационный объект и инструментальное средство обработки, или в более общем случае – информационную и логическую концепции проектирования ЗОИС.

Таким образом, анализ процесса проектирования ИИС с общих гносеологических позиций позволяет выделить следующие этапы этого вида научно-познавательной деятельности [9, 12]. Прежде всего, это выделение некоторой части объективной реальности (предметной области), с моделью которой необходимо работать. На следующем этапе происходит выявление наиболее существенных черт, явлений ПдО, определение элементарных понятий и их взаимосвязей, законов и ограничений, управляющих развитием данной ПдО, – т.е. создается некоторый образ реальности, отражающий (пока еще) нестрогое ее восприятие исследователем. Для фиксации приобретаемых знаний в виде текста в рамках некоторой знаковой системы необходимо определить формализм, который позволил бы это сделать. Собственно, необходим язык, по возможности более строгий, чем естественный, и в то же время не слишком отличающийся по восприятию от последнего. Необходим также метод (методы), позволяющий осуществить формализацию знаний. Таким может быть сочетание логико-информационного и аксиоматического методов. Поэтому следующий этап процесса проектирования заключается в построении формальной теории исследуемой ПдО в соответствии с принципами формализации. Построенная таким образом формальная теория подлежит интерпретации, т.е. каждому выражению теории нужно задать некоторым образом соответствие действительному положению вещей, иначе, каждому понятию сопоставить конкретный объект реального мира, а каждому суждению – связь реальных объектов, соответствующих входящим в данное суждение понятиям, и т.п. После этого можно определять истинность или ложность высказываний теории, т.е. уже можно судить о правильности рассуждений о ПдО. В результате этого можно исследовать построенную модель в строгом научном смысле, ставить по отношению к последней вопросы о ее полноте, непротиворечивости, разрешимости, прослеживать ее эволюцию и другие моменты и, наконец, интерпретировать полученные в результате решения этих задач знания в заданной ПдО.

Цепочка информационных технологий «Компьютерная обработка ЕЯ (*Natural Language Processing* (*NLP*))→Представление знаний (*Knowledge Representation* (*KR*))→Обработка знаний (*Knowledge Processing* (*KP*))» представляет собой реализацию базовых процедур анализа, синтеза и понимания естественного языка (ЕЯ) компьютером, которые в более широком смысле можно выразить продукционной цепочкой: *входное_сообщение→система_знаний→реакция*. Суть этой цепочки определяется интеграцией лингвистических и предметных знаний, что в общем случае представляет интегрированную информационную технологию в стадии интенсивного развития исследований (в том числе и фундаментальных) как по центральной проблеме – представления знаний и методов их использования, – так и по принципам, методам, подходам, технологии построения интеллектуальных систем, разработке их архитектуры и технологии применения (рис. 5). Здесь ЛКТ – лингвистический корпус текстов.

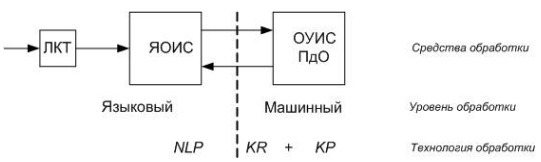

Рис. 5. Схема обработки знаний, содержащихся в ЕЯО

Задачи и особенности проектирования ОУИС

В известных методах формализованного проектирования информационных систем [1] собственно процесс проектирования представля-

ется в виде последовательности этапов (в основном системного, алгоритмического и логического), на каждом из которых проект представлен совокупностью математических моделей, описывающих различные ее части. Указанная совокупность математических моделей тесно связана с системой взаимосвязанных алгоритмов, которые, в свою очередь, описывают соответствующее множество решаемых задач и в своей совокупности представляют общий алгоритм проектирования ИС.

Применительно к проектированию ОУИС обобщенная последовательность решаемых задач анализа и синтеза на всех этапах следующая [22, 23].

1. Постановка задачи:

- исследование заданной предметной области,
- анализ класса решаемых задач в заданной ПдО,
- выбор основных критериев проектирования ОУИС.

2. Разработка инфологической модели ОУИС:

- разработка информационной модели ОУИС,
- разработка онтологической модели языковых знаний (языково-онтологической картины мира),
- разработка онтологической модели предметных знаний (онтологии ПдО),
- разработка модели системной интеграции языковых и предметных знаний,
- разработка модели формально-логического описания языковых и предметных знаний.

3. Разработка системы взаимосвязанных алгоритмов функционирования ОУИС:

- разработка алгоритмов функционирования ЯОИС,
- разработка алгоритмов функционирования ОУИС ПдО,
- разработка алгоритмов перехода от обработки ЕЯО к обработке предметных знаний.

4. Разработка архитектуры и структуры ОУИС:

- разработка архитектурно-структурной организации ЯОИС (разработка знание-ориентированного лингвистического процессора, разработка базы знаний лексики ЕЯ),
- разработка архитектурно-структурной организации ОУИС ПдО (онтологической базы знаний ПдО, машины вывода, системы прикладного процессинга, семантической памяти),
- разработка интерфейса пользователя.

5. Проверка функционирования ОУИС в соответствии с заданными критериями проектирования.

Особенности архитектуры и структуры онтологических или «онтологизированных» знание-ориентированных систем полностью определяют подход к их проектированию. Исходя из приведенного анализа, эти особенности можно разделить на системные и технологические, языковые и проблемные.

Системные особенности вытекают из требования построения на базе ОУИС инструментального комплекса обработки знаний в заданной предметной области с заданными техническими характеристиками, в том числе с возможностью подключения аппаратных средств поддержки, спроектированных на базе современных ПЛИС-технологий, высокими потребительскими свойствами.

Технологические особенности определяются состоянием развития элементно-технологической базы современных компьютеров и связаны, прежде всего, с появлением сверхмощных программируемых логических интегральных схем (ПЛИС), вызвавших интенсивное развитие новых методов и средств проектирования ЗОИС.

Архитектурно-структурная организация ЗОИС

Современные архитектуры КС ориентированы на обработку знаний. В 80-х годах прошлого столетия японскими специалистами для преодоления семантического разрыва между человеком и компьютером был предложен ряд архитектур КС, отвечающих требованиям обработки знаний под общим названием «архитектура ЭВМ пятого поколения» [26] (однако поставленная конечная цель перед проектом так и не была достигнута). Анализ указанных архитектур показал, что их основными компонентами были, кроме машин баз знаний, и ма-

шины обработки структурированных данных. А для машин обработки неструктурированных данных, в частности ЕЯО, исследования по созданию новых архитектур не проводились. Очевидно, это было связанно с тем, что (как указывалось выше) принципиально невозможно полностью формализовать ЕЯО. Справедливости ради следует отметить, что ряд идей, выдвинутых в процессе реализации проекта по созданию ЭВМ пятого поколения, в новом качестве могут быть использованы и в настоящее время для архитектур обработки ЕЯ, в частности применение интеллектуальной многопортовой памяти, сортировки, параллельной обработки синтаксических единиц ЕЯО и др.

Новым толчком в активизации исследований по созданию новых архитектур (и соответствующих структур) машин обработки ЕЯ послужило начатое в 90-х годах новое направление, названное онтологическим инжинирингом, как развитие *Knowledge engineering*. С этим направлением и в настоящее время связывают успехи в разработке новых архитектур КС обработки знаний.

Анализ когнитивного процесса познания, основных источников знаний, известных технологий обработки языковых и предметных знаний, архитектурных особенностей информационных систем, реализующих указанные технологии (что в некоторой степени зафиксировано на рис. 5), позволяет сделать некоторые, общего характера выводы относительно архитектурно-структурной организации современной ОУИС. В частности:

– архитектуру ОУИС целесообразно строить на базе двух, относительно независимых информационных подсистем – ЯОИС и ОУИС ПдО;

– необходимы формализованные средства (и их интерпретатор), позволяющие осуществить переход от обработки текстовой информации на основе языковых знаний к обработке формализованных предметных знаний;

– экспертные знания, представленные в виде естественно-языковых спецификаций, могут эффективно обрабатываться в такой ОУИС;

– необходимо интегрировать в единую три, относительно независимые ИТ – *NLP*, *KR* и *KP*.

Достаточно подробное описание архитектурно-структурной организации ЯОИС приведено в [1, 9, 21–23]. Далее приведём описание системы обработки предметных знаний – ОУИС ПдО, а также общие замечания в части архитектурно-структурной организации ОУИС с онтолого-управляемой архитектурой (рис. 6).

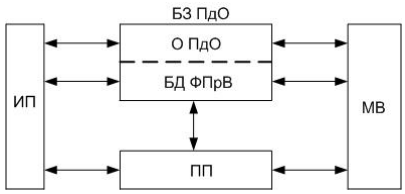

Рис. 6. Блок-схема ОУИС ПдО

Компоненты ОУИС ПдО

На основе анализа известных положений об архитектурно-структурных особенностях и потребительских признаках ЗОИС [1, 2, 11], определим состав и назначение основных компонентов онтолого-управляемой информационной системы обработки знаний в произвольной предметной области (ОУИС ПдО).

ОУИС ПдО содержит онтологическую базу знаний (в общем случае конечное множество системно интегрированных онтологических баз знаний), машину вывода, систему прикладного процессинга и интерфейс с пользователями и/или внешней средой. Компоненты ОУИС ПдО характеризуются совокупностью следующих признаков:

• *онтологическая БЗ* является компьютерной формой представления модели ПдО конечного объема. Она состоит из концептуальной надстройки (иерархической структуры концептуальных понятий) и фактографической базы, включающей в себя базу фактов и базу правил вывода новых фактов. В свою очередь, концептуальная надстройка имеет связи (или связана концептуальными отношениями) с понятиями более высокого уровня абстракции (входящих в так называемую метаонтологию домена прикладных областей, например медицина, материаловедение, право и др.);

• *машина вывода* (МВ) реализует внутренний алгоритм интерпретации структур знаний применительно к входным данным, взаимодействует с системой прикладного процессинга (ПП), принимает от него входные данные и выдает полученный в процессе интерпретации результат;

• *система ПП* осуществляет в диалоге с пользователем и/или внешней средой следующие функции: принимает задания и формирует обобщенный алгоритм их реализации в виде последовательности процедур; формирует задания для МВ и получает от него результаты; выдает результаты решения задач через интерфейс пользователя (ИП); осуществляет ввод, редактирование и пополнение знаний в БЗ ПдО;

• *интерфейс пользователя* осуществляет перевод данных заданий/результатов в их внутреннее (машинное)/внешнее (язык пользователя) представление.

Одним из главных назначений такой ОУИС (прежде всего с точки зрения пользователя) является то, что в ней должна быть зафиксирована вся релевантная информация об интересующей его (пользователя) части реального мира, чтобы иметь возможность манипулировать данной информацией и моделировать рассуждения человека о правилах, законах и ограничениях, действующих в этой части реального мира (ПдО) и управляющих ее развитием, о других ее свойствах и характерных особенностях.

Архитектурно-структурная организация ОУИС

Архитектурно-структурная организация онтолого-управляемой информационной системы обработки знаний, содержащихся в ЕЯО, представлена на рис. 7.

Здесь приняты следующие обозначения: ЗОПС – знание-ориентированная поисковая система, ЯПЗ – язык представления знаний, ЛБД – лексикографическая база данных, ЛО ПдО – лингвистическая онтология предметной области.

Отметим следующие особенности архитектуры:

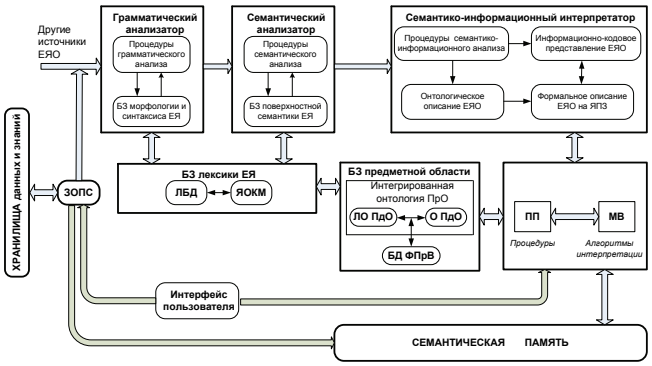

Рис. 7. Архитектурно-структурная организация ОУИС

► ЗОПС для найденных множеств текстовых документов строит (совместно с ЯОИС) соответствующие им онтологии, сохраняемые в семантической памяти. При этом поиск необходимой информации по онтологии текстового документа (семантическому графу) существенно упрощается при расширении процедур обработки текста;

► интерфейс служит для распознавания запросов пользователей, составленных в виде естественно-языковых предложений;

► семантическая память в общем случае ориентирована на широкий спектр хранимой информации и настраивается на круг задач прикладного процессинга. Можно выделить следующие, наиболее важные сегменты:

• библиотека онтологий прикладного домена;

• метаонтология прикладного домена;

• корпоративная память, ЭБ или коллекция документов, хранящихся в виде онтологий;

• память прикладного процессинга;

• некоторые дополнительные возможности, описанные в [21, 24].

Заключение. Анализ совокупности задач, включающий в себя онтологический, логический и методологический аспекты проектирования знание-ориентированных информационных систем, позволяет говорить о наличии научно-технической проблемы, заключающейся в отсутствии проверенных на практике теоретических наработок и эффективных программно-аппаратных систем для компьютерной обработки знаний, содержащихся в ЕЯО. Место онтологических (языковых и предметных) знаний,

формализованное описание и использование их в информационной системе является центральной идеей в разрабатываемой методологии.

Полученные результаты позволят приблизиться к решению таких проблем, как организация и использование больших баз формализованных знаний и создание концептуально-методологических и технологических основ теории междисциплинарных научных исследований.

1. *Палагин А.В.* Архитектура онтологоуправляемых компьютерных систем // Кибернетика и системный анализ. – 2006. – № 2. – С. 111–124.
2. *Кургаев А.Ф.* Проблемная ориентация архитектуры компьютерных систем. – К.: Сталь, 2008. – 540 с.
3. *Ракитов А.И.* Курс лекций по логике науки. – М.: Высшая школа, 1971. – 176 с.
4. *Кондаков Н.И.* Логический словарь-справочник. – М.: Наука, 1975. – 720 с.
5. *Гаазе-Рапопорт М.Г.*, *Поспелов Д.А.* Структура исследований в области искусственного интеллекта // Толковый словарь по искусственному интеллекту. – М.: Радио и связь, 1992. – С. 5–20.
6. *Налимов В.В.* Спонтанность сознания: вероятностная теория смыслов и смысловая архитектоника личности. – М.: Прометей, 1989. – 288 с.
7. *Хилькевич А.П.* Проблема расширения традиционной силлогистики. – Минск: Изд. БГУ, 1981. – 115 с.
8. *Рыков В.В.* Управление знаниями. – http://ryk-kypc2.narod.ru/part2.doc
9. *Палагин А.В., Яковлев Ю.С.* Системная интеграция средств компьютерной техники. – Винница: УНІВЕРСУМ, 2005. – 680 с.
10. *Гаврилова Т.А.*, *Хорошевский В.Ф.* Базы знаний интеллектуальных систем. – СПб.: Питер, 2001. – 384 с.
11. *Кургаев А.Ф.* Анализ развития идеала структуры научной теории // Кибернетика и вычислительная техника. – 2003. – Вып. 139. – С. 50–63.
12. *Андон Ф.И.*, *Яшунин Л.Е.*, *Резниченко В.И.* Логические модели интеллектуальных информационных систем. – К.: Наук. думка, 1999. – 397 с.
13. *Люггер Дж.Ф.* Искусственный интеллект. Стратегии и методы решения сложных проблем. – М.: Изд. дом «Вильямс», 2003. – 864 с.
14. *Башмаков А.И.*, *Башмаков И.А.* Интеллектуальные информационные технологии: Учеб. пособие. – М.: Изд-во МГТУ им. Н.Э. Баумана, 2005. – 304 с.
15. *Искусственный* интеллект: В 3 кн. Системы общения и экспертные системы: Справочник / Под ред. *Э.В. Попова*. – М.: Радио и связь, 1990. – Кн. 1. – 464 с.
16. *Искусственный* интеллект. – В 3-х кн. Модели и методы: Справочник / Под ред. *Д.А. Поспелова*. – М.: Наука, 1990. – Кн. 2. – 304 с.
17. *Искусственный* интеллект: В 3 кн. Программные и аппаратные средства: Справочник / Под ред. *В.Н.* Захарова, В.Ф. Хорошевского. – М.: Радио и связь, 1990. – Кн. 3. – 368 с.
18. *Гладун В.П.* Процессы формирования новых знаний. – София: СД «Педагог 6», 1994. – 192 с.
19. *Кривий С.Л.* Дискретна математика: Вибр. питання: Навч. посіб. для студ. вищ. навч. закл. – К.: Вид. дім. «Києво-Могилянська академія», 2007. – 572 с.
20. *Палагин А.В.* К решению основной задачи эмуляции // УСиМ. – 1980. – № 3. – С. 24–28.
21. *Палагін О.В.*, *Петренко М.Г.* Розбудова абстрактної моделі мовно-онтологічної інформаційної системи // Математичні машини і системи. – 2007. – № 1. – С. 42–50.
22. *Палагин А.В.*, *Петренко Н.Г.* К проектированию онтологоуправляемой информационной системы с обработкой естественно-языковых объектов // Там же. – 2008. – № 2. – С. 14–23.
23. *Палагин А.В.*, *Петренко Н.Г.* К вопросу проектирования онтологоуправляемой ИС обработки ЕЯО. International Book Series «INFORMATION SCIENCE & COMPUTING». – Varna, Bulgaria. – 2008. – N 2. – P. 160–164.
24. *Sowa*, *John F.* Knowledge Representation: Logical, Philosophical, and Computational Foundations, Brooks Cole Publ. Co., Pacific Grove, CA, © 2000. – 594 p.
25. *Рубашкин В.Ш.* Представление и анализ смысла в интеллектуальных информационных системах. – М.: Наука, 1989. – 191 с.
26. *ЭВМ* пятого поколения: Концепции, проблемы, перспективы / Под ред. *Т. Мотоока*; Пер. с англ.; Предисл. Е.П. Велихова. – М.: Финансы и статистика, 1984. – 110 с.

Поступила 27.02.2009
Тел. для справок: (044) 526-3328 (Киев)