

ВВЕДЕНИЕ

В настоящее время интеллектуальные информационные технологии (ИИТ) всё глубже проникают во все сферы научно-технической деятельности человеческого общества. Центральным их звеном являются технологии инженерии знаний, включая процессы управления знаниями, поэтому успехи данного направления во многом определяются интеллектуальным уровнем и общей эффективностью компьютерных систем (КС). К сожалению, многие идеи этого раздела искусственного интеллекта (ИИ) сегодня так и остаются не реализованными полностью или частично по ряду причин.

Первой причиной является недостаточное понимание исследователями принципов и механизмов восприятия, отображения, осознания и обработки информации/знаний в мозге человека и, как следствие, соответствующие компьютерные методы и алгоритмы не совершенны. Не вызывает сомнения тот факт, что ИИТ и соответствующие КС будут и дальше 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 совокупного смысла, заложенного в анализируемых знаковых системах. При этом под смыслом понимается то, что делает знаковую систему текстом, а последний – источником знаний.

Второй причиной является недостаточная эффективность большинства современных КС: последние в процессе управления знаниями должны оперировать не примитивными данными (в традиционном понимании), а знаниями, представленными в подходящей формальной теории. Разработка и использование новых информационных технологий, таких как суперкомпьютерные и GRID-вычисления, многоагентные системы и др., со своей стороны, также требуют знание-ориентированного подхода. Ориентация на аппаратную поддержку (реализацию) основных процедур манипулирования знаниями на базе современных микроэлектронных технологий (например, ПЛИС-технологии, PIM-технологии, реконфигурируемого компьютеринга) с применением соответствующих САПР поможет в значительной степени повысить эффективность современных КС.

Особенностью интеллектуальных информационных систем (ИИС) являются интеграционные процессы, которые влияют на развитие их архитектуры и функциональных возможностей. Системная интеграция междисциплинарных знаний как таковых, а также технологии их эффективного формирования, представления, обработки и использования являются сильными стимулирующими факторами на этом пути. Не менее важным фактором является естественная интеграция двух областей ИИ, которые раньше развивались почти параллельно и

независимо: knowledge-engineering и компьютерная лингвистика. Указанную интеграцию можно рассматривать как закономерное эволюционное развитие разных научных теорий, заложенных ещё с давних времён и известных как логика и онтология. Отсутствие развитой теоретической базы системологии междисциплинарного взаимодействия и квалифицированных кадров (инженер по знаниям должен быть специалистом высокой квалификации в областях онтологического инжиниринга, формальных логических теорий, прикладной лингвистики, целой совокупности математических теорий, когнитологии, психологии и др.) является, по-видимому, *третьей причиной*.

Сегодня очевидно, что развитие онтологического инжиниринга, а также знание-ориентированных информационных систем (ЗОИС) связано с совершенствованием методов компьютерной обработки ЕЯ-информации и предметных знаний. Нетрудно заметить, что хорошо проработанные методы и средства обработки ЕЯ-информации для решения прикладных задач в узкоспециализированных предметных областях (ПдО) недостаточны для решения комплексных задач, связанных с формально-логическим представлением ЕЯ-информации, извлечением предметных знаний и их последующим использованием в произвольных ПдО.

Отсюда следует актуальность и важность разработки новых научных методов и подходов к автоматическому анализу и глубинно-семантическому пониманию ЕЯ-информации, её формализованному представлению в рамках подходящей теории, автоматическому обнаружению и извлечению новых знаний, технологий и, наконец, соответствующих инструментальных средств автоматизированного построения баз знаний предметных областей.

Решение данной проблемы является основополагающим на пути разработки общей теории понимания, формирования, представления и обработки естественно-языковых знаний методами ИИ и является целью данной работы.

Исследования этой проблемы традиционны для направления инженерии знаний. Большое внимание ей уделял ещё В.М. Глушков, в частности сформулировав и описав концептуальную парадигму работы со знаниями. Известны весомые вклады в её решение Н. Хомским, М. Минским, А. Тарским, Д. Дэвидсоном, Монтегю, Д.А. Поспеловым, Т.А. Гавриловой, Н.Н. Леонтьевой, Э.В. Поповым, В.А. Широковым и другими. Многие из работ по другим направлениям ИИ, математической логике можно также отнести к данной области.

Особенностью данных исследований является их междисциплинарность, сочетание различных теорий, решений и средств их достижения. И вместе с тем, как указывалось выше, единой теории пока нет.

Структурно весь материал монографии представлен в виде восьми логически связанных глав и четырёх приложений.

В *первой главе* описан общий подход к исследованию предметно-ориентированных знаний, начиная с онто-гносеологического анализа. Такой анализ рассматривает абстрактную систему <Субъект – Объект – Отношение> и взаимодействие её составных частей между собой. Методы изучения каждого такого взаимодействия рассматриваются в разных научных теориях – гносеологии, онтологии, теории отражения и логике.

Рассмотрена естественная схема взаимодействия человека с окружающим миром: *входная информация* → “*сознание – понимание – знания*”. В ней выделено компонент “сознание”, играющий роль персонифицированного инструмента, который вырабатывает совокупность предметно, ситуативно или причинно-связанных сущностей, которые образуют “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 языковой статус, и это выражено понятием “языковое сознание”, а сознательная картина мира при этом представлена “языковой” картиной мира (ЯКМ), которая является не только системой знаний, а и некоторым буфером, который связывает общие знания с профессиональными. Все знания человечества (или полная картина мира) представлены в виде “пирамиды” знаний, состоящей из частичных картин мира. При этом предполагается, что каждая картина мира строится по тем же законам системного подхода, что и ЯКМ, а интерпретация процесса развития знаний может быть использована как в рамках одной предметной области, так и для описания полной картины мира.

Выполнен сущностный анализ понятия “знание” как *исходной, преобразующей и конечной* составляющей процесса познания (“знание посредством знаний преобразуется в новое знание”). Предложены частные определения таких взаимосвязанных понятий, как *смысл, знание, данные и информация*, выражающие собой наибольшую проблемность исследований в области построения систем обработки знаний, в том числе и ЗОИС, оперирующих ЕЯ-информацией. Рассмотрены основные свойства знаний – внутренняя интерпретируемость, структурированность, связность, семантическая метрика и активность. На основе выполненного анализа синтезирована обобщённая схема “эволюционирования” знаний.

Используя известные представления *онтологий верхнего уровня*, разработаны методологические основы построения многоуровневых компьютерных онтологий и системного анализа, необходимого для построения строгих математических моделей произвольной предметной области (проблемы, задачи) и их интерпретации для получения конкретных результатов, и в частности *компьютерной онтологии верхнего уровня и системы категорий* как знаний об общей структуре действительности. При этом система категорий выступает по

отношению ко всякому новому знанию как некоторый фиксированный *категориальный каркас*, который имеется до познания данного конкретного класса явлений и определяет формы этого познания, так что всякое новое содержание, доставляемое конкретным исследователем, подводится под одну из категорий и упорядочивается по правилам категориального каркаса *Универсального*, рассматриваемого как три (относительно) независимые группы категорий: *Материальное – Абстрактное; Независимое – Относительное – Промежуточное; Продолжительное – Событийное*. Для категорий онтологии верхнего уровня предложена формально описанная система аксиом.

Рассмотрена формальная постановка задачи извлечения знаний из естественно-языковых текстов (ЕЯТ), основным моментом в которой является построение двух фундаментальных отношений, присутствующих практически в каждом ЕЯТ. Это отношения эквивалентности и частичного порядка, которые названы *общеуказанными*. Первое отношение определяет классы синонимических объектов, а второе отношение – иерархию подчинённости классов эквивалентности. Оба эти отношения составляют основу построения онтологий.

Выполнен анализ трёх базовых аспектов разработки онтолого-управляемых информационных систем, а именно – *онтологический, логический и методологический*. Эти аспекты имеют свои, в общем случае фиксированные объекты исследований, соответствующие процессу познания или разработке некоторой ИИС. В связи с ними все объекты могут быть поделены на три группы: система сущностей (или объектов реального мира), система знаний и система обработки сущностей в соответствии с данной системой знаний. Первая группа является предметом онтологического исследования, вторая группа – логического исследования и третья – методологического.

Разработана онтолого-классификационная схема концептуальных понятий, отображающая проблематику (структуру) многочисленных научных исследований в области ИИ, нижние уровни которой представляют объект и предмет исследований при решении научно-технической проблемы построения онтолого-управляемой информационной системы (ОУИС) с обработкой больших объёмов ЕЯ-информации.

Проанализированы известные структурные направления онтологического инжиниринга как перспективного инструмента предметных и междисциплинарных научных исследований и предложены их расширения, основными из которых являются системная интеграция научных знаний (включая персонифицированные знания) и инструментальный комплекс онтологического назначения (ИКОН), “работающий” на протяжении всего процесса реализации интегрированной информационной технологии обработки ЕЯ-информации.

Во *второй главе* рассмотрены логико-математические основы, используемые при разработке и проектировании ОУИС. Основное внимание уделено теориям множеств, отношений и алгоритмов. При этом рассмотрены формальные логические языки, такие как исчисление высказываний и исчисление предикатов первого порядка. Приведены многочисленные примеры, поясняющие использование рассматриваемых математических понятий.

Читателям, знакомым с основами указанных математических теорий, эту главу можно пропустить без ущерба понимания остального материала.

В *третьей главе* рассмотрены различные формы представления знаний, извлечённых из естественно-языковых высказываний (ЕЯВ), для подмножеств логики первого порядка, а также некоторые формализованные аспекты извлечения новых знаний. При этом описание сопровождается многочисленными примерами.

Выполнен сравнительный анализ формальных логических языков первого порядка для представления ЕЯВ, таких как силлогизмы Аристотеля, логика высказываний и алгебра логики, исчисление предикатов, экзистенциально-конъюнктивная логика, продукционная система, дизъюнкты Хорна, фреймовые системы и объектно-ориентированные системы, семантические сети, концептуальные графы (КГ) и язык KIF (Knowledge Interchange Format).

На основе выполненного анализа (по существенным характеристикам) подмножеств FOL (First Order Logic), в том числе с их реализациями в действующих системах и нотациях, сделан вывод, что в них можно представить сложные утвердительные предложения и вопросы к их частям, заменяя некоторые компоненты представления (силлогизма, формулы, графа) неизвестной переменной, ответы на которые могут быть получены или решением уравнения, где значение формул с переменной должно быть истинным, или поиском фрагмента графа-вопроса в общем графе и установлением значений неизвестных узлов или дуг. Результаты анализа представлены в виде таблицы, из которой видно, что: наибольшей выразительностью обладают три системы: исчисление предикатов, концептуальные графы и KIF. Это может служить критерием выбора именно этих систем логики в качестве основных для представления и анализа знаний. Выбор формы представления и анализа знаний в общем случае зависит от конкретной задачи. Некоторые задачи можно эффективно решать навигацией в КГ по алгоритмам работы с графами. Ряд задач не имеет методов и средств их решения на графах – в таком случае следует применять универсальные общие методы логического вывода в системе исчисления предикатов или KIF.

Детально рассмотрено формализованное представление ЕЯВ в форме КГ, даны неформальные определения КГ и семантической сети, приведены примеры, поясняющие эти понятия. Рассмотрено также представление ЕЯВ в форме концептуальных графов Дж. Совы, вы-

явлено ряд недостатков такого представления. Предложены пути устранения этих недостатков, которые основаны на принципах и механизмах построения языково-онтологической информационной системы (ЯОИС) с задействованием таких компонент как “*Языково-онтологическая картина мира (ЯОКМ)*”, “*Онтология ПдО*” и “*Грамматический процессор*”.

Раскрыты методы извлечения знаний из ЕЯТ на примерах лексико-грамматического анализа, силлогистики Аристотеля и её теоретико-множественной интерпретации.

В *четвёртой главе* рассмотрены принципы организации и методика построения *компьютерных онтологий*. Сформулировано определение компьютерной онтологии и онтологической системы, состоящей из *онтологии объектов, онтологии процессов и онтологии задач*, а также общую методологию проектирования онтологических знаний ПдО.

Особый статус предметных знаний приобретают языковые знания и соответствующие онтологии. Рассмотрены особенности лингвистических онтологий и проанализирована наиболее известная из них – WordNet. Основные характеристики наиболее известных онтологий сведены в таблицу.

Рассмотрена ЯОКМ – лингвистическая онтология, которая ориентируется на украинский и русский языки и состоит из лексикографической базы данных, моделей морфологических, синтаксических, поверхностно-семантических и глубинно-семантических структур ЕЯ, модели интегрированной интерпретации входного ЕЯТ.

Выделен особый вид системной интеграции междисциплинарных знаний – интеграции лингвистических и предметных знаний. Проанализированы известные проекты такой интеграции и предложена формализация её некоторых аспектов и оценок.

Выполнен анализ известных методик IDEF (Integrated DEFinition) – проектирования сложных объектов, в том числе онтологий, и сравнение с предложенным *системно-онтологическим анализом*. Оказывается, что методики IDEF и системно-онтологический подход используют одни и те же множества сущностей ПдО и проблемного пространства (ПрП), акцентируя внимание на различных совокупностях характеристик и атрибутов; конечные цели этих двух подходов отличаются: для IDEF – функционально-блочная модель, а для системно-онтологического подхода – онтолого-содержательная модель.

Описаны методики проектирования онтологии объектов, онтологии процессов, онтологии задач и соответствующий обобщённый алгоритм, при этом предпочтение отдано автоматизированному способу проектирования, так как даже для сравнительно небольших предметных областей списки терминов-понятий, терминов-процессов, терминов-отношений и терминов, описывающих классы задач, могут составлять несколько тысяч записей.

В *пятой главе* рассмотрены оригинальные алгебро-логические и лингвистические модели формализации и обработки ЕЯТ, а также формальная постановка задачи анализа. Выделены особенности анализа ЕЯТ, определяющие направленность на формирование понятийной структуры, т. е. на автоматическое извлечение знаний из текстов и их прагматическую интерпретацию в терминах прикладных задач. При этом текст рассматривается как объект разных уровней анализа: как знаковая система, как грамматическая система, система знаний о мире, источник исходных данных для решения прикладных задач. Каждый уровень имеет свои особенности, свои средства выражения, а, следовательно, предполагает наличие специфических методов обработки.

Предложен алгебро-логический подход к анализу и пониманию ЕЯТ, включающий алгебру и систему списочных структур. Введено понятие *списка элементов* и определены функции и операции над ними, а в качестве элементов списка рассматриваются слова естественного языка (ЕЯ). При этом алгебраическая система получается путём добавления в алгебру списочных структур *предиката равенства* и *условного оператора*.

Предложена информационная модель этапов лингвистического анализа и рассмотрено их функционирование на содержательном уровне.

В *шестой главе* разработаны методологические основы проектирования ЗОИС с обработкой естественно-языковых объектов (ЕЯО) на основе онтологического подхода. Такие ЗОИС в монографии названы онтолого-управляемыми (или с онтолого-управляемой архитектурой). В онтолого-управляемую информационную систему входят две подсистемы: языково-онтологическая информационная подсистема, обеспечивающая обработку языковых знаний, и онтолого-управляемая информационная подсистема (ОУИС ПДО), обеспечивающая обработку знаний некоторой предметной области.

Рассмотрены основы *онто-логического подхода* проектирования ОУИС. Такой подход учитывает логико-информационную и онтологическую концепции проектирования, а также виртуальную парадигму, при которой архитектура компьютерной системы ориентирована на технологию реконфигурируемого процессинга. Последняя обеспечивает адаптивность системы благодаря наличию в ОУИС архитектурных и технологических возможностей настройки в условиях априорной и текущей неопределённостей на основе обучения и опыта. Важной отличительной особенностью онто-логического подхода по сравнению с логико-информационным (кроме онтологизации) является добавление ещё одного (предметного) уровня управления ОУИС.

Предложена блок-схема процесса синтеза структуры ОУИС, состоящая из блоков синтеза ЯОИС, ОУИС ПДО и отображения языковых знаний в их формальное представление.

Сформулированы характерные черты ЗОИС с онтолого-управляемой архитектурой, из которых основными являются: композиция онтологий разного уровня и назначения, как по вертикали, так и по горизонтали (по вертикали интегрируются онтология верхнего уровня, онтология домена предметных областей и онтология предметной области, а по горизонтали интегрируются онтологии предметных знаний и знаний проблемного пространства). Особо отмечено, что предметные знания взаимодействуют с языковыми знаниями через лингвистическую онтологию предметной области; что эффективно многократное использование онтологий предметной области и онтологий задач для разных наборов типовых задач, а вследствие двойной парадигмы предметных знаний и онтологического управления оказывается возможным включение компонентов архитектуры ОУИС в архитектурно-структурную организацию инструментального онтологического комплекса автоматизированного построения онтологических баз знаний предметных областей.

Рассмотрена архитектура и информационная модель ЯОИС, одной из отличительных особенностей которой в сравнении с известными лингвистическими процессорами является наличие ЯОКМ. Последняя, по сути, представляет собой онтологическую базу данных лексики естественного языка, компьютерно-ориентированную информационную структуру, которая является необходимым компонентом любой знание-ориентированной системы, обрабатывающей ЕЯТ.

Рассмотрена архитектурно-структурная организация ОУИС ПдО, содержащей онтологическую базу знаний (в общем случае конечное множество системно интегрированных онтологических баз знаний), машину вывода, подсистему прикладного процессинга и интерфейс с пользователями и/или внешней средой, а также обобщённый алгоритм проектирования ОУИС в целом.

Седьмая глава содержит описание инструментального комплекса онтологического назначения, который реализует относительно самостоятельные интегральные технологии – *автоматизированного построения онтологии ПдО* на основе обработки соответствующего лингвистического корпуса текстов и *системной интеграции онтологических знаний* в некотором домене предметных областей.

Разработана методология проектирования ИКОН, алгоритмов его функционирования и программно-аппаратных средств. Для этого проанализированы функциональные возможности известных инструментальных средств (ИнС) построения онтологий с учётом таких факторов как поддерживаемые формализмы и формат представления, множество предоставляемых пользователю сценариев работы с онтологическими структурами, архитектура приложения, интерфейс пользователя и доступность. Выявлен ряд недостатков, основные из кото-

рых: отсутствие процедур автоматического (автоматизированного) формирования компонент онтологии; англоязычный интерфейс с пользователем, в котором (для большинства ИнС) не предусмотрено присвоение имён компонентам онтологии на русском или украинском языке; структуризация концептов в онтологии выполняется только по одному типу отношений; для большинства общедоступных ИнС не предусмотрена работа с большими объёмами онтологических знаний.

ИКОН состоит из трёх подсистем, интегрирующих информационные ресурсы, программно-аппаратные средства обработки и естественный интеллект инженера по знаниям и эксперта, которые, взаимодействуя между собой, реализуют совокупность алгоритмов автоматизированного итерационного построения понятийных структур предметных знаний, их накопления и/или системной интеграции. Программно-аппаратные средства ИКОН реализованы по модульному принципу. Разработаны обобщённые алгоритмы проектирования.

Представлены методологические основы системной интеграции междисциплинарных научных знаний, разработаны алгоритм и программные средства объединения двух и более онтологий в близких предметных областях.

Восьмая глава посвящена прикладным аспектам функционирования ОУИС и инструментального комплекса онтологического назначения. Рассмотрены примеры автоматизированного построения онтологической базы знаний и онтологии задач ПдО “Проектирование вычислительных устройств и систем на ПЛИС”, применения онтологического подхода к разработке электронных курсов обучения, в частности для предметной дисциплины “Базы данных”, формализованного построения онтологии текстового документа и разработки некоторых процедур Text Processing. Предложена общая архитектура знание-ориентированной поисковой системы, функционирующей в составе ИКОН и ОУИС.

В **приложении** приведены: описания онтологий верхнего уровня SUMO и Mikrokosmos; методика проектирования простой онтологии ПдО; онтограф домена “Информатика” и краткий толковый словарь по онтологическому инжинирингу.

Изложенный в монографии материал относится к *Computer Science*, в частности, разработке знание-ориентированных информационных систем, компьютерных онтологий и формализации процесса анализа и понимания естественно-языковых текстов.