

На правах рукописи



Васильев Иван Анатольевич

МЕТОДЫ И ИНСТРУМЕНТАЛЬНЫЕ СРЕДСТВА
ПОСТРОЕНИЯ СЕМАНТИЧЕСКИХ WEB-ПОРТАЛОВ

Специальность: 05.13.11 – Математическое и программное обеспечение вычислительных машин, комплексов и компьютерных сетей

АВТОРЕФЕРАТ
диссертации на соискание ученой степени
кандидата технических наук

Томск – 2005

Работа выполнена в Томском политехническом университете

Научный руководитель: доктор технических наук,
профессор Ямпольский Владимир Захарович

Официальные оппоненты: доктор технических наук,
профессор Бондаренко Владимир Петрович

кандидат технических наук,
с.н.с. Ходашинский Илья Александрович

Ведущая организация: Институт систем информатики имени А. П. Ершова СО РАН
(г. Новосибирск)

Защита состоится «14» декабря 2005 г. в 16:30 на заседании диссертационного совета Д 212.269.06 при Томском политехническом университете по адресу: г. Томск, ул. Советская, 84.

Заверенные отзывы на автореферат просьба направлять в двух экземплярах по адресу: 634004, г. Томск, ул. Советская, 84, ученому секретарю диссертационного совета Д 212.269.06.

С диссертацией можно ознакомиться в библиотеке Томского политехнического университета.

Автореферат разослан «___» _____ 2005 г.

Ученый секретарь
диссертационного совета,
кандидат технических наук



Сонькин М. А.

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность темы. Совершенствование существующих и разработка новых подходов к сбору, хранению, обработке и распространению информации является неотъемлемой частью процесса развития информационных технологий и информационных систем (ИС). Необходимость такого совершенствования во многом обусловлена непрерывным ростом объема электронных документов и их доступности, что наряду со слабой структурированностью информационных фондов осложняет работу пользователей с ними. Существующие подходы к работе с информацией становятся не достаточно эффективными.

Для решения проблемы совершенствования доступа к растущему объему информации и информационным услугам, предоставляемым многочисленными источниками информации, была предложена концепция порталов. Портал является программной системой, которая призвана обеспечить унифицированный доступ к информации, хранящейся во множестве разнородных информационных источников. Портал структурирует информацию и предоставляет средства для ее поиска.

Различные виды порталов разрабатываются и внедряются в России и за рубежом. Перспективность данного подхода к интеграции и структуризации информации отмечается аналитиками и подтверждается пользователями. Однако большие объемы информации, интегрируемые порталами, поставили вопрос о повышении эффективности и качества обработки информации в порталах.

Одним из подходов к решению данной проблемы является использование семантических технологий, призванных обеспечить обработку информации в информационных системах с учетом ее семантики.

Рассматриваемые в данном диссертационном исследовании порталы являются многопользовательскими ИС, которые предоставляют унифицированный доступ к различным информационным источникам и программным приложениям. Порталы, как правило, обрабатывают большой объем информации. С учетом этого применение в рамках порталов новых подходов и методов к обработке информации имеет высокую практическую значимость, а исследование подходов и разработка методов построения порталов на основе семантических технологий являются актуальными.

Цель работы – *разработка методов использования семантических технологий в порталах для реализации информационных процессов в них с учетом семантики контента объектов.*

Для достижения поставленной цели исследования **необходимо решить следующие задачи:**

- разработать архитектуру семантического ядра портала;
- разработать методы семантического описания контента объектов портала;
- разработать методы использования описаний объектов портала для реализации его функций на семантическом уровне.

Объектом исследования являются технологии построения порталов.

Предметом исследования являются подходы и методы использования семантических технологий в порталах для реализации информационных процессов на семантическом уровне.

Методы исследования. В ходе диссертационного исследования были использованы модели и методы теории множеств, профессионально-логический анализ и обобщение, метод экспертных оценок, методы объектно-ориентированного проектирования и программирования.

Научная новизна результатов исследования заключается в следующем:

- разработан метод семантического описания объектов портала с точки зрения контента, использующий предложенную автором структуру семантических метаданных;

- разработан метод вычисления семантической близости метаданных, основанный на известном методе определения ситипности;
- разработаны методы поиска, категоризации и формирования рекомендаций объектов портала с учетом семантики их контента, основанные на методе вычисления близости семантических метаданных;
- разработана архитектура семантического ядра портала, реализующего функции описания семантики контента объектов, поиска, категоризации и предоставления рекомендаций.

Практическая значимость исследования заключается:

- в программной реализации разработанного автором семантического ядра портала;
- в применении предложенных методов для разработки семантического портала для современной IT-компания;
- в возможности использования созданного семантического ядра портала в системах управления знаниями.

Реализация и внедрение результатов работы. Разработанное семантическое ядро портала, основывающееся на предложенных методах, полностью программно реализовано на языке C#. Вклад автора в программную реализацию составил 88% (более 15 тысяч строк кода). Результаты исследования использовались при разработке и реализации портала для «Центра профессиональной подготовки специалистов нефтегазового дела» ТПУ (г. Томск) и портала системы управления знаниями компании «ЭлеСи» (г. Томск).

Апробация работы. Основные положения и отдельные результаты работы докладывались и обсуждались на следующих конференциях и семинарах:

- Международная научно-практическая конференция студентов, аспирантов и молодых ученых «Современные техника и технологии 2003».
- Международная научно-практическая конференция студентов, аспирантов и молодых ученых «Современные техника и технологии 2004».
- Международная научно-практическая конференция «Современные средства и системы автоматизации 2004».

Публикации. По теме диссертации опубликовано 9 научных статей, в том числе одна в рецензируемом издании. Имеется свидетельство государственного координационного центра информационных технологий об отраслевой регистрации разработки «Web-портал для работы с явными и неявным знаниями организации» в Отраслевом фонде алгоритмов и программ (свидетельство №4608; авторы Тузовский А. Ф., Васильев И. А., Козлов С. В., Усов М. В.; дата выдачи 29.04.2005).

Личный вклад автора. Все результаты, составляющие основное содержание диссертации, получены автором самостоятельно. В опубликованных работах лично автором обоснованы варианты использования семантических технологий в информационных системах в общем [1] и в частности в порталах [2, 3, 4], пояснены разработанные методы описания семантики объектов портала и вычисления их семантической близости [5, 6], описано разработанное семантическое ядро портала [5] и приведены варианты применения разработанных методов и алгоритмов в работе порталов [5, 7, 8, 9].

Структура диссертации. Диссертация состоит из введения, 4 глав, заключения, списка литературных источников из 117 наименований и 6 приложений. Содержит 56 рисунков и 36 таблиц.

ОСНОВНОЕ СОДЕРЖАНИЕ РАБОТЫ

Во введении обоснована актуальность темы диссертационной работы, сформулированы цель и задачи исследования, приведено краткое содержание работы.

В первой главе описаны результаты анализа исследований, выполненных отечественными и зарубежными авторами в области создания традиционных и семантических порталов.

Рассмотрена классификация порталов по различным критериям и показано, что как определения понятия «портал», так и функциональные возможности его программных реализаций варьируются.

На основании анализа описаний, опубликованных компаниями-производителями традиционных порталов, а также на основании анализа литературы выявлены наиболее часто встречающиеся функции портала. В качестве отличительных функций портала обозначены: интеграция информационных источников, интеграция приложений и поиск по всем информационным источникам. Предложена структура портала, включающая инфраструктуру и функциональные модули. Выявлено, что подходы, технологии и стандарты, используемые как в рамках инфраструктуры портала, так и при реализации его функциональных модулей, являются общеупотребительными. Это позволило описать обобщенную архитектуру портала.

Хотя подходы к разработке и реализации порталов можно считать достаточно проработанными с точки зрения методов и используемых технологий, существует объективная необходимость их развития. Эта необходимость обусловлена развитием телекоммуникационных технологий, делающих информацию более доступной, а также объективным ростом объема информации, что в свою очередь требует повышения качества информационных процессов.

В качестве подхода к решению указанной проблемы в исследовании рассматриваются активно развивающиеся в настоящее время семантические технологии. В рамках семантических технологий разрабатываются подходы, стандарты и методы, которые обеспечивают возможность явного представления семантики информации. Явное представление семантики информации должно способствовать созданию программных систем, позволяющих обрабатывать информацию на семантическом уровне. В зависимости от области применения и решаемой задачи подходы к представлению семантики и ее обработке могут варьироваться, но их объединяет наличие модели знаний, которая описывает семантику отдельных элементов информации и связи между ними.

На сегодняшний день в рамках семантических технологий наиболее активно исследуется и развивается онтологический подход к представлению знаний предметной области, на основании которого разрабатываются интеллектуальные информационные системы, и в том числе порталы.

Онтология – модель представления знаний какой-либо предметной области в виде набора понятий этой предметной области и существующих между ними отношений. То есть онтология представляет предметную область в виде некоторой сетевой структуры, в которой семантика каждого понятия определяется через его отношения с другими понятиями. Причем во множестве отношений существует отношение типа «родитель-ребенок», упорядочивающее понятия предметной области в иерархию – таксономию понятий. К отношениям того типа относятся отношения «целое-часть» (part-of), «класс-подкласс» (is-a) и т.п. В диссертации приведен обзор типов онтологий и языков описания онтологий.

Онтологическая модель предметной области задает семантику понятий, которые используются для описания информационных объектов портала. Такие описания называются семантическими метаданными, и они позволяют: 1) устранить лексическую многозначность терминов, используемых для описания информационных объектов; 2) определять соответствие между различными информационными объектами, используя онтологию. Семантические метаданные могут описывать объект с разных точек зрения: структуры, контекста и контента. Описание контента наиболее важно для тех ИС, которые реализуют функции полнотекстовой обработки информации. В диссертации рассмотрены существующие структуры и языки описания семантических метаданных и сделан вывод о

том, что исследованиям семантических метаданных, описывающих контент, не уделяется должного внимания. Данная проблема сдерживает практическую реализацию полнотекстовой обработки информации с учетом ее семантики. В порталах, одним из основных объектов которых являются электронные документы, эта проблема особенно актуальна.

Порталы, использующих семантические технологии для реализации информационных процессов, называются семантическими порталами (СП). В ходе диссертационного исследования выполнен анализ существующих семантических порталов. В результате анализа необходимо констатировать следующее:

- исследования в области семантических технологий и их внедрения в порталы в настоящее время активно ведутся (в основном за рубежом);
- подходы к использованию семантических технологий в порталах варьируются;
- в области семантических технологий наработано множество различных программных инструментов, и стандартов;
- методам описания контента информационных объектов портала не уделяется достаточного внимания.

Поэтому целью диссертационного исследования является разработка методов, позволяющих использовать семантические технологии в порталах для реализации информационных процессов в них с учетом семантики контента объектов.

Во второй главе обосновывается архитектура семантического ядра, являющегося центральной подсистемой СП и предоставляющего функции, позволяющие обрабатывать объекты портала на семантическом уровне.

Проведен анализ и обобщение возможных вариантов использования онтологий. Для реализации информационных процессов в портале с учетом семантики объектов предложены варианты использования онтологий:

1. описание объектов портала;
2. семантический поиск;
3. формирование списка объектов, связанных с исходным объектом;
4. формирование списка объектов, похожих на исходный объект.

Обосновано использование дескриптивной логики (ДЛ) класса *SHIQ* в качестве логического формализма представления онтологической модели, а языка *OWL DL* – в качестве языка записи онтологии для ее хранения и передачи. На основании выбранных средств представления знаний дано формальное определение онтологии O_{DL} , используемое для описания предлагаемой структуры семантических метаданных и предлагаемых методов по работе с семантикой объектов портала.

Определение 1. Онтология – это знаковая система

$$O_{DL} = \langle C, CD, R, A, I, V, R_I, A_I, L, P_C, P_R, P_A, P_{IC}, P_{LC}, P_{LR}, P_{LA}, P_{LI} \rangle, \text{ где} \quad (1)$$

$C = \{c_1, \dots, c_n\}$ – конечное множество понятий в онтологии,

$CD = \{cd_1, \dots, cd_l\}$ – множество стандартных типов данных, включающее два типа $\{string, integer\}$,

$R = \{r_1, \dots, r_m\}$ – конечное множество бинарных отношений $r_i(c_x, c_y)$ между понятиями,

$A = \{a_1, \dots, a_w\}$ – конечное множество атрибутов, т.е. бинарных отношений $a_i(c_x, cd_y)$ между понятиями и стандартными типами данных,

$I = \{i_1, \dots, i_l\}$ – конечное множество экземпляров в онтологии,

$V = \{v_1, \dots, v_q\}$ – конечное множество конкретных значений стандартного типа,

$R_I = \{ri_1, \dots, ri_m\}$ – конечное множество конкретизированных отношений, т.е. бинарных отношений $ri_i(i_x, i_y)$ между экземплярами,

$A_I = \{ai_1, \dots, ai_w\}$ – конечное множество конкретизированных атрибутов, т.е. бинарных отношений $ai_i(i_x, v_y)$ между экземпляром и конкретными значениями,

$L = \{l_1, \dots, l_k\}$ – конечное множество лексических меток (словарь онтологии),

$P_C \subseteq C \times C$, $P_C \in R$ – антисимметричное, транзитивное, нерефлексивное бинарное отношение, являющееся отношением частичного порядка на множестве понятий C ,

$P_R \subseteq R \times R$ – антисимметричное, транзитивное, нерефлексивное бинарное отношение, являющееся отношением частичного порядка на множестве отношений R ,

$P_A \subseteq A \times A$ – антисимметричное, транзитивное, нерефлексивное бинарное отношение, являющееся отношением частичного порядка на множестве атрибутов A ,

$P_{IC} \subseteq I \times C$ – бинарное отношение инцидентности между множествами I и C ,

$P_{LC} \subseteq L \times C$ – бинарное отношение инцидентности между множествами L и C ,

$P_{LR} \subseteq L \times R$ – бинарное отношение инцидентности между множествами L и R ,

$P_{LA} \subseteq L \times A$ – бинарное отношение инцидентности между множествами L и A ,

$P_{LI} \subseteq L \times I$ – бинарное отношение инцидентности между множествами L и I .

В соответствии с выбранным подходом к определению онтологии предложена структура семантических метаданных, позволяющая описывать объекты портала с использованием не только отдельных элементов онтологии (понятий и экземпляров), но и с использованием триплетов элементов (таблица 1).

Таблица 1. Значения семантического атрибута

№	Возможные значения		
	Субъект	Предикат	Объект
1	c_x		
2	i_x		
3	c_x	r_y	c_z
4	c_x	r_y	i_z
5	i_x	r_y	c_z
6	i_x	r_y	i_z
7	c_x	a_y	v_z
8	i_x	a_y	v_z

В таблице 1 значения 1 и 2 представляют собой единичные элементы из онтологии, а значения 3-8 являются триплетами.

Семантические метаданные «связывают» объекты портала с частями предметной области. Исходя из того, что «строительными блоками» онтологии являются триплеты (семантика понятия определяется его отношениями с другими понятиями), то и использование триплетов в семантических метаданных позволяет наиболее точно отражать взаимосвязь объекта портала с частями предметной области. Ясно, что с помощью триплетов невозможно выразить семантику контента объекта абсолютно точно. Триплеты не учитывают временных, модальных и прочих характеристик описываемой информации. Для этого нужны более сложные конструкции наподобие схем концептуального анализа Шенка. Но у семантических метаданных другая функция. Они описывают семантику объекта с точки зрения контента, основываясь на онтологии. Следовательно, семантические метаданные не могут описать контент объекта точнее, чем

это позволяет онтология. Чем точнее описана предметная область в виде онтологии, тем точнее можно описать контент объектов портала с использованием этой онтологии.

С учетом предложенной структуры метаданных и определения онтологии O_{DL} дано формальное определение семантических метаданных MD_{DL} .

Определение 2. Семантические метаданные с возможностью использования триплетов – это множество

$$MD_{DL}(q_i) = MD_{SN}(q_i) \cup MD_{TR}(q_i), \text{ где} \quad (2)$$

$Q = \{q_1, \dots, q_k\}$ – конечное множество объектов семантического портала,

$MD_{SN}(q_i) = \{(sn_{i1}, k_{i1}), \dots, (sn_{in}, k_{in})\}$ – множество отдельных элементов онтологии, описывающих объект q_i ,

$sn_{in} \in C \cup I$ – отдельный элемент онтологии,

$k_{in} \in (0;1]$ – коэффициент, обозначающий релевантность отдельного элемента sn_{in} объекту q_i ,

$MD_{TR}(q_i) = \{(tr_{i1}, k_{i1}), \dots, (tr_{if}, k_{if})\}$ – множество триплетов, описывающих объект q_i ,

$tr_{if} = \langle subj_{if}, pred_{if}, obj_{if} \rangle$ – триплет

$subj_{if} \in C \cup I$ – субъект в триплете

$pred_{if} \in R \cup A$ – предикат в триплете

$obj_{if} \in C \cup I \cup V$ – объект в триплете

$k_{if} \in (0;1]$ – коэффициент, обозначающий релевантность триплета tr_{if} объекту q_i .

Описанный подход к определению и использованию онтологий и семантических метаданных позволил разработать методы формирования описаний объектов портала и их обработки с учетом семантики контента. Учет семантики контента объектов не подразумевает их детального семантического анализа. Переход на семантический уровень осуществляется за счет устранения синтаксической многозначности описаний и учета связей между понятиями предметной области. То есть разработанные методы предоставляют следующие возможности:

1. Учет в процессе обработки информационных объектов наличия омонимии и полисемии в естественном языке. Это достигается за счет моделирования знаний предметной области в виде онтологии, которая содержит понятия, которые в свою очередь имеют множественные лексические представления. В результате появляется возможность выявления омонимов и многозначных слов в предметной области и возможность устранения неоднозначности.
2. Учет в процессе обработки информационных объектов наличия эквивалентных лексических конструкций (синонимов) в естественном языке. Это достигается за счет закрепления за понятиями онтологии множественных лексических представлений – синонимов. В результате появляется возможность сравнения синтаксически различной, но семантически похожей информации.
3. Учет в процессе обработки информационных объектов иерархической природы понятий, используемых для описания объектов. Это достигается за счет использования таксономии понятий.

Для предоставления возможности использования разработанных методов в семантических порталах разработана архитектура семантического ядра (рис. 1), включающая сервер онтологий и сервер семантических метаданных.

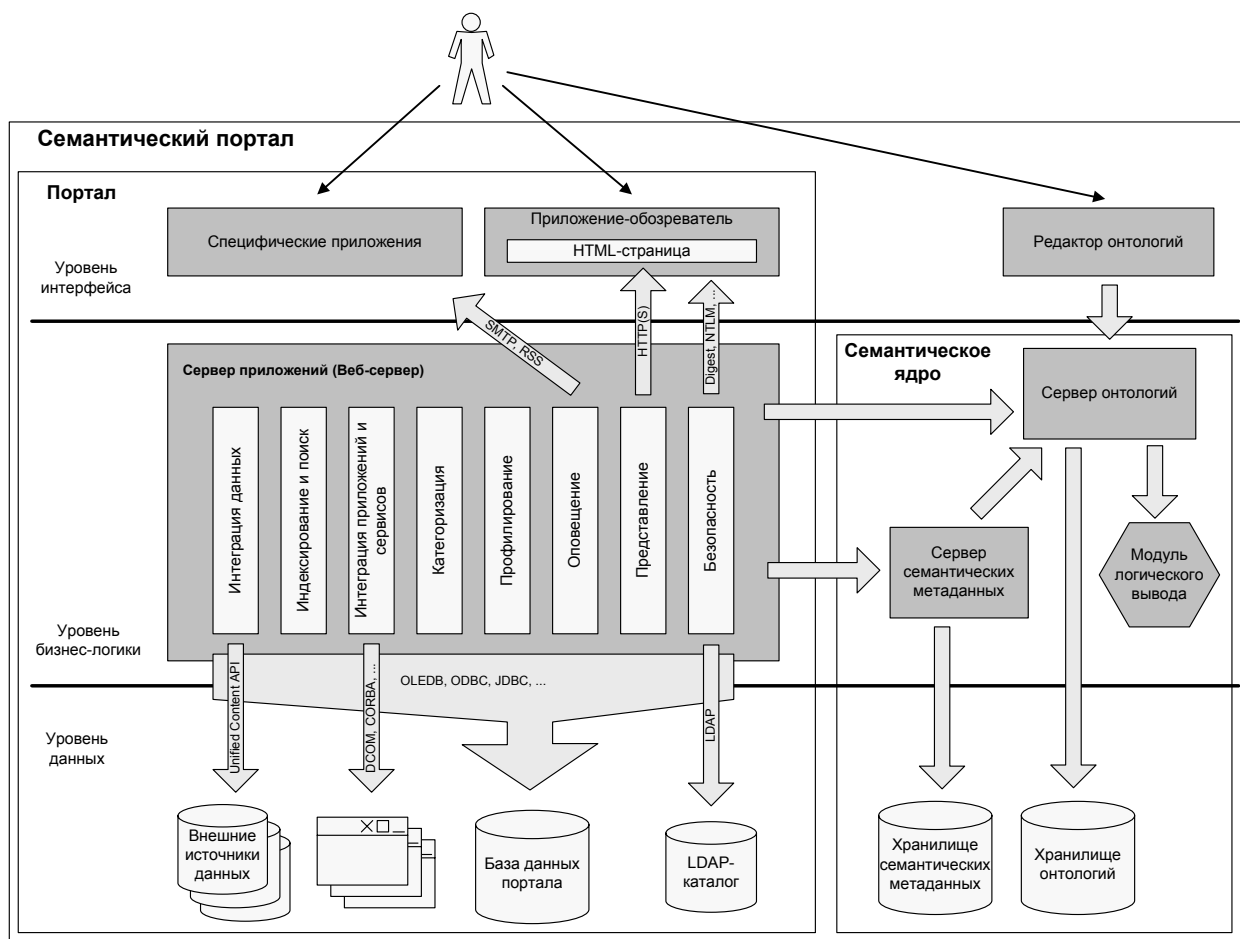


Рис. 1. Место семантического ядра в структуре портала

Сервер онтологий (СО) – это отдельно функционирующая программная система, хранящая множество онтологий и предоставляющая к ним доступ. Для реализации в семантическом ядре выбранных вариантов использования онтологии сервер онтологий предоставляет следующие функции: 1) хранение онтологий; 2) предоставление онтологий; 3) логический вывод; 4) поиск в онтологии запрашиваемых понятий и отношений.

Сервер семантических метаданных (ССМ) – это отдельно функционирующая программная система, хранящая семантические метаданные, предоставляющая к ним доступ и обрабатывающая их. Для реализации выбранных вариантов использования онтологии сервер семантических метаданных, тесно взаимодействуя с сервером онтологий, предоставляет следующие функции: 1) формирование семантических метаданных; 2) хранение семантических метаданных; 3) предоставление семантических метаданных; 4) сравнение семантических метаданных.

В третьей главе поясняются разработанные методы формирования описаний объектов портала и их обработки с учетом семантики контента.

Обосновано, что онтологическая модель, используемая семантическим ядром портала, должна иметь структуру, в которой выделена неизменная часть – онтология приложения, и переменная часть, которая включает онтологии верхнего уровня, предметных областей и задач.

Разработанный метод формирования семантических метаданных определяет правила выбора элементов онтологии для описания объектов, а также определяет алгоритм поиска понятий и экземпляров в тексте. Формирование семантических метаданных объекта портала должен выполнять человек. Он должен в соответствии с сущностью объекта описания определять элементы семантических метаданных. Элементы представляют собой либо триплеты со структурой «субъект–предикат–объект», либо

отдельные понятия или экземпляры из онтологии, задающие «субъект». Если субъект указывается человеком таким образом, чтобы отражать сущность предмета описания, то на выбор предиката и объекта накладываются дополнительные ограничения. Множество возможных предикатов в триплете ограничивается выбранным субъектом триплета. В таблице 2 приведены правила формирования множества M_{PRED} возможных предикатов в триплете.

Таблица 2. Правила определения возможного значения предиката в триплете

Значение субъекта	Правило
Понятие c_x	$M_{PRED} = \{pr_i \in R \cup A \mid pr_i(c_x, c_y) \vee pr_i(c_x, cd_y)\}$
Экземпляр i_x	$C_{INST}(i_x) = \{c_i \in C \mid P_{IC}(i_x, c_i)\}$ $M_{PRED} = \{pr_i \in R \cup A \mid (pr_i(c_x, c_y) \vee pr_i(c_x, cd_y)) \wedge c_x \in C_{INST}(i_x)\}$

Множество возможных объектов зависит от выбранного предиката. Правила формирования множества M_{OBJ} возможных объектов в триплете приведены в таблице 3.

Таблица 3. Правила определения возможного значения объекта в триплете

Значение предиката	Правило
Отношение r_x	$M_{OBJ} = \{obj_i \in C \cup I \mid r_x(c_x, obj_i) \vee (r_x(c_x, c_y) \wedge P_{IC}(obj_i, c_y))\}$
Атрибут a_x	$M_{OBJ} = \{obj_i \in cd_j \mid a_x(c_x, cd_j)\}$

Если семантические метаданные формируются на основании текстового описания объекта, то в дополнение к правилам выбора предикатов и объектов используется алгоритм поиска понятий и экземпляров в тексте. Это позволяет частично автоматизировать процесс выбора субъекта из онтологии. С этой целью текстовое описание анализируется на наличие понятий и экземпляров, которые могут выступать в качестве субъектов в элементах семантических метаданных. Результатом работы алгоритма являются множество понятий M_C и множество экземпляров M_I , которые могут иметь отношение к объекту, для которого формируются семантические метаданные.

Разработанный метод вычисления семантической близости элементов онтологии основан на определении онтологии O_{DL} . Он развивает существующий подход к оценке близости объектов, основанный на использовании сотипности (cotopy). Сотипность объектов – это оценка схожести положения сравниваемых объектов в некоторой иерархии.

$$S(c_k, c_l) = \frac{|C_{ANC}(c_k) \cap C_{ANC}(c_l)|}{|C_{ANC}(c_k) \cup C_{ANC}(c_l)|} - \text{сотипность, где} \quad (3)$$

$$C_{ANC}(c_i) = \{c_j \in C \mid P_C(c_j, c_i) \vee c_j = c_i\} \quad (4)$$

На основании метода вычисления сотипности разработаны методы вычисления семантической близости элементов онтологии (таблица 4). Пара сравниваемых элементов рассматривается как упорядоченная в том смысле, что первый элемент пары является эталоном, с которым сравнивается второй элемент пары – кандидат. Из этого следует, что в общем случае показатель семантической близости упорядоченной пары элементов (O_1 , O_2) может быть не равным показателю семантической близости упорядоченной пары элементов (O_2 , O_1).

Таблица 4. Допустимые сравнения между элементами онтологии

Кандидат Эталон	Понятие	Экземпляр	Отношение	Атрибут	Значение
Понятие	+	+			
Экземпляр	+	+			
Отношение			+		
Атрибут				+	
Значение					+

Разработанный метод вычисления близости семантических метаданных основан на определении онтологии O_{DL} , определении семантических метаданных MD_{DL} и использует метод вычисления семантической близости элементов онтологии. Для оценки близости семантических метаданных предложены три показателя:

1. Показатель семантической близости пересекающихся метаданных без учета наследования

$$SM_{FO}(MD_{DL}(q_i), MD_{DL}(q_j)) = \frac{\sum_{smd_{ix} \in MD_{DL}(q_i)} \max_{smd_{jy} \in MD_{DL}(q_j)}^m (k_{ix} * k_{jy} * SE_F(smd_{ix}, smd_{jy}))}{n} \quad (5)$$

2. Показатель семантической близости пересекающихся метаданных с учетом наследования

$$SM_{CO}(MD_{DL}(q_i), MD_{DL}(q_j)) = \frac{\sum_{smd_{ix} \in MD_{DL}(q_i)} \max_{smd_{jy} \in MD_{DL}(q_j)}^m (k_{ix} * k_{jy} * SE_C(smd_{ix}, smd_{jy}))}{n} \quad (6)$$

3. Показатель семантической близости перекрывающихся метаданных с учетом наследования

$$SM_{CS}(MD_{DL}(q_i), MD_{DL}(q_j)) = \begin{cases} SM_{CO}(MD_{DL}(q_i), MD_{DL}(q_j)), & \text{если } \prod \max(SE_C(smd_{ix}, smd_{jy})) > 0 \\ 0, & \text{иначе} \end{cases} \quad (7)$$

При расчете данных показателей определяется близости элементов метаданных без учета наследования $SE_F(smd_{ix}, smd_{jy})$ и с учетом наследования $SE_C(smd_{ix}, smd_{jy})$ с использованием метода вычисления семантической близости элементов онтологии.

Метод вычисления близости семантических метаданных позволяет количественно оценить схожесть между двумя объектами. На практике оценка близости обычно выполняется между объектом-эталон и множеством объектов-кандидатов. Например, поисковый запрос (эталон) сравнивается с описаниями документов (множество кандидатов), хранящихся в портале, в результате чего формируется множество релевантных запросу документов. То есть метод вычисления близости семантических метаданных можно рассматривать как средство ранжирования объектов-кандидатов на основании объекта-эталона. После ранжирования те кандидаты, семантическая близость которых эталону меньше некоторого порогового значения, исключаются из результирующего множества объектов.

Очевидно, что чем больше множество кандидатов, тем дольше выполняется ранжирование. Поэтому уменьшение множества кандидатов за счет исключения из него объектов с заведомо низким показателем семантической близости способно увеличить вычислительную эффективность операции семантического сравнения. С этой целью

разработан метод фильтрации множества кандидатов. Метод фильтрации основан на использовании системы логического вывода для дескриптивной логики. Он позволяет отфильтровать из множества кандидатов те объекты, семантическая близость которых объекту-эталону равна нулю.

Метод вычисления близости семантических метаданных применяется в комплексе с методом вычисления семантической близости элементов онтологии и методом фильтрации для реализации в семантических порталах функций поиска, категоризации и формирования рекомендаций. Для выполнения семантического поиска используется показатель $SM_{CS}(MD_{DL}(q_i), MD_{DL}(q_j))$, для категоризации – $SM_{CO}(MD_{DL}(q_i), MD_{DL}(q_j))$, а для формирования рекомендаций – $SM_{FO}(MD_{DL}(q_i), MD_{DL}(q_j))$.

В четвертой главе описывается программная реализация и внедрение разработанных структур, методов и алгоритмов построения семантических Web-порталов.

Выполнено проектирование сервера онтологий и сервера семантических метаданных, составляющих семантическое ядро портала.

Сервер онтологий реализует (рис. 2) объектную модель онтологии, описанной на языке OWL DL, и позволяет:

- получать доступ к функциям сервера онтологий с использованием технологии .NET Remoting;
- хранить файлы с описаниями онтологий в файловой системе;
- выполнять запросы к онтологии, использующие логический вывод.

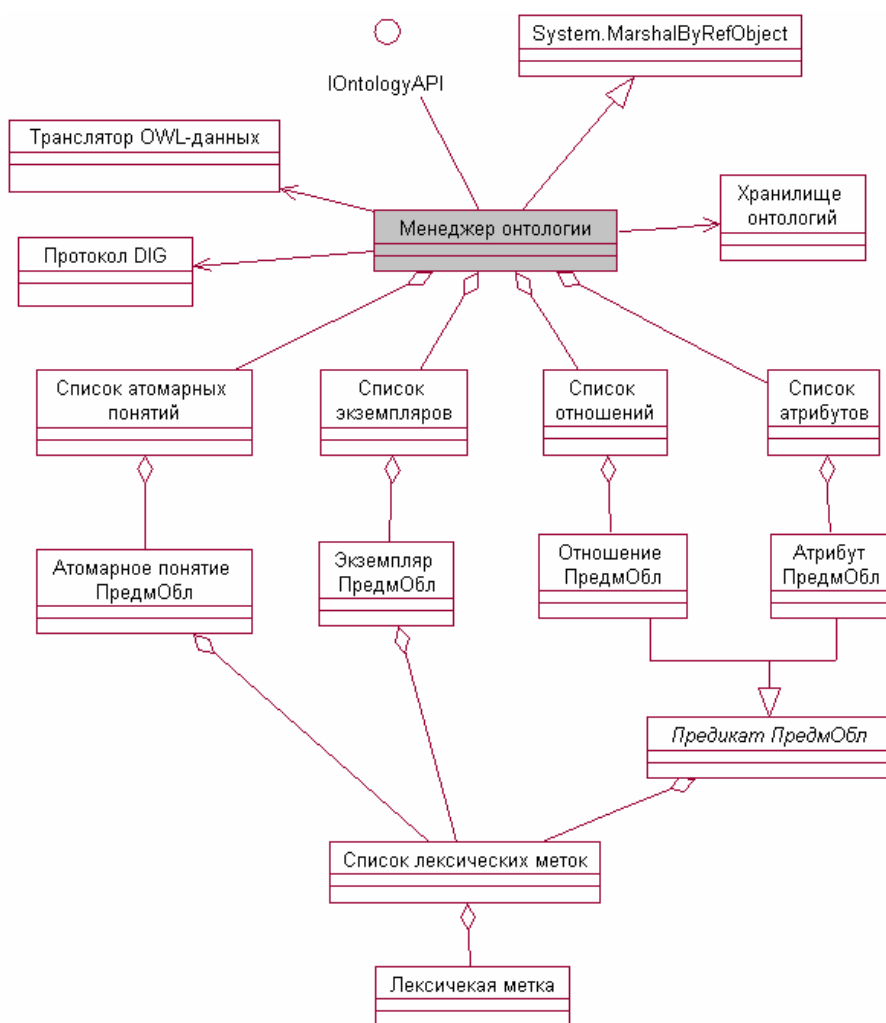


Рис. 2. UML-диаграмма классов, реализуемых CO

Объектная модель онтологии содержит классы-сущности, представляющие элементы онтологии, класс для трансляции OWL-данных во внутреннее представление сервера и управляющий класс, реализующий программный интерфейс доступа к функциям сервера онтологий.

В качестве классов-сущностей были выделены: атомарное понятие, экземпляр, отношение, атрибут, целочисленное значение и строковое значение. Для первых четырех типов сущностей в онтологии заданы лексические метки, а для двух оставшихся типов лексическими метками являются их значения, представленные в текстовом виде. Лексические метки составляют словарь онтологии.

Класс «Транслятор OWL-данных» предназначен для трансляции описания онтологии на языке OWL DL в систему взаимосвязанных объектов соответствующих классов-сущностей. Результатом трансляции является наборы выявленных понятий, экземпляров, атрибутов, отношений, строковых и целочисленных значений. Сгенерированное внутреннее объектное представление онтологии используется для ускорения доступа к элементам онтологии и выполнения запросов.

Управляющий класс «Менеджер онтологии» реализует программный интерфейс *IOntologyAPI*, определяющий методы и свойства сервера онтологий. Для использования технологии .NET Remoting управляющий класс «Менеджер онтологии» помимо реализации интерфейса *IOntologyAPI* наследует функциональность класса *System.MarshalByRefObject*, входящего в состав инфраструктуры Microsoft .NET.

В качестве подсистемы хранения онтологий, описанных на языке OWL DL, была выбрана файловая система. Такой подход достаточно прост для реализации и одновременно оставляет возможность использования существующих редакторов онтологий. Все редакторы онтологий поддерживают работу с файловой системой. Для работы с файловой системой менеджер онтологии использует класс «Хранилище онтологий».

В качестве системы логического вывода (СЛВ) для дескриптивной логики была выбрана свободно распространяемая система RACER версии 1.7.24. Система RACER реализует логический вывод для дескриптивной логики класса $ALCQHI_{R+}(D^-)$, которая расширяет атрибутивный язык (*AL*) такими возможностями как произвольное отрицание (*C*), транзитивные отношения (R_+), инверсные отношения (*I*), иерархия отношений (*H*), количественные ограничения на отношения (*Q*) и некоторые конкретные домены (D^-). В качестве конкретных доменов поддерживаются строки и числа. Дескриптивная логика класса $ALCQHI_{R+}(D^-)$ по выразительности является подклассом *SHIQ*, не поддерживая лишь перечисляемые типы (*nominals*). На сегодняшний день система RACER реализует наиболее выразительную дескриптивную логику с использованием высокопроизводительного алгоритма (*tableau-based algorithm*) логического вывода, который используется для обработки онтологий, описанных на языке OWL DL.

Для взаимодействия с СЛВ RACER был выбран протокол DIG, так как он был специально разработан для взаимодействия с различными СЛВ, основанными на ДЛ. Использование этого протокола устраняет зависимость от RACER и при появлении новых СЛВ позволяет использовать их.

Второй сервер – сервер семантических метаданных реализует (рис. 3) объектную модель семантических метаданных предложенной структуры и позволяет:

- получать доступ к функциям сервера семантических метаданных с использованием технологии .NET Remoting;
- хранить семантические метаданные различных объектов в реляционной базе данных Microsoft SQL Server 2000;
- вычислять близость семантических метаданных с использованием предложенного метода.

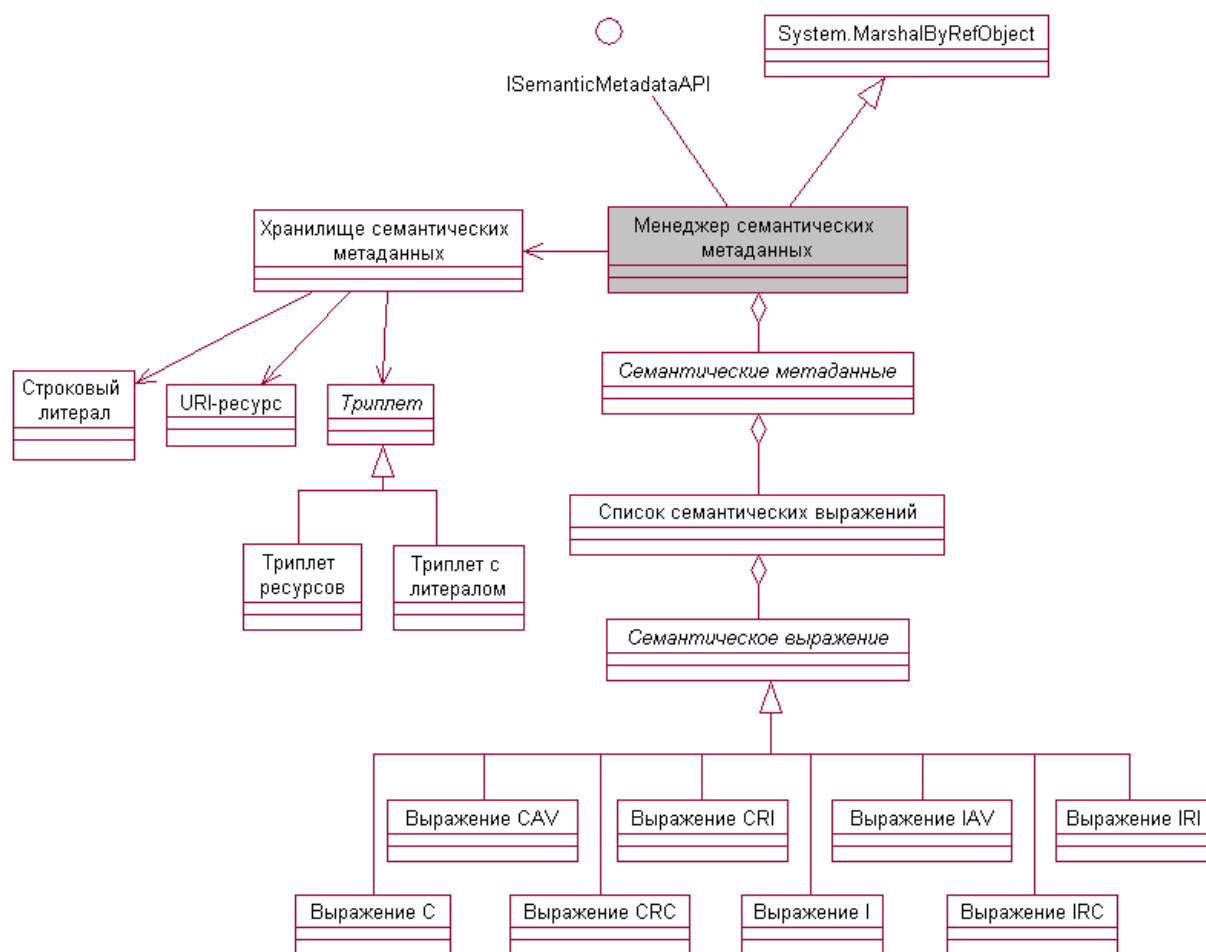


Рис. 3. UML-диаграмма классов, реализуемых CCM

Объектная модель семантических метаданных включает классы-сущности для элементов языка RDF, сам класс семантических метаданных, классы семантических выражений, входящих в состав семантических метаданных, и управляющий класс.

В соответствии со спецификацией языка RDF его основными понятиями являются ресурс, литерал и триплет. Триплет может быть двух типов: «ресурс – ресурс – ресурс» и «ресурс – ресурс – литерал». Для целей обмена данными между хранилищем семантических метаданных и управляющим классом были разработаны соответствующие классы-сущности элементов языка RDF.

Семантические метаданные представлены классом, объекты которого хранят множество семантических выражений, а также предоставляют вспомогательные методы. Количество семантических выражений в метаданных неограниченно. Для каждого из восьми типов семантических выражений (таблица 1) разработан соответствующий класс.

Управляющий класс «Менеджер семантических метаданных» реализует программный интерфейс ISemanticMetadataAPI, который определяет методы и свойства сервера семантических метаданных.

Сервер семантических метаданных предоставляет доступ к реализуемым свойствам и методам посредством технологии .NET Remoting. Для этого реализующий его класс – менеджер семантических метаданных – наследуется от системного класса System.MarshalByRefObject.

Для хранения семантических метаданных разработано специальное хранилище на базе СУБД Microsoft SQL Server 2000. Данное хранилище RDF-данных отличается от существующих решений (например, RDFGateway, Sesame). Во-первых, оно не предоставляет возможность выполнения запросов. Это обусловлено тем, что знания

описываются на языке OWL DL, более выразительном, чем RDF, и поэтому функции запросов и логического вывода реализованы в сервере онтологий. Во-вторых, хранилище разработано с возможностью группировки RDF-данных. В существующих решениях все RDF-высказывания, относящиеся к одной онтологии, хранятся совместно. В разработанном хранилище есть возможность группировки RDF-высказываний в семантические метаданные отдельных объектов описания. Такой подход позволяет отделить общие знания от знаний об отдельных объектах описания. Это в свою очередь позволяет реализовать функцию семантического поиска объектов портала, а не элементов онтологии. ССМ получает доступ к хранилищу посредством объекта класса «Хранилище семантических метаданных», который предоставляет методы по добавлению, изменению, получению и удалению RDF-данных.

Реализация предложенного метода вычисления близости семантических метаданных выполнена в виде трех функций класса «Менеджер семантических метаданных», каждая из которых позволяет вычислить соответствующий показатель близости (выражения 5, 6, 7). Для расчета этих показателей используется сервер онтологий, предоставляющий функции логического вывода и запросов к онтологии. В частности онтология используется для расчета семантической близости элементов онтологии, являющихся частью семантических метаданных.

Разработанное и спроектированное семантическое ядро портала, включающее сервер онтологий и сервер семантических метаданных, полностью реализовано программно на платформе Microsoft .NET с использованием языка программирования C#. В общей сложности для программной реализации было разработано с использованием языка моделирования UML 177 классов и интерфейсов, а объем кода составил более 16 тысяч строк. Вклад автора диссертационного исследования в разработку и проектирование составляет 100%, а в программную реализацию – 88%.

Тестирование функций семантического ядра выполнялось с использованием онтологии, созданной для части предметной области «Автоматизация технологических процессов». Общее количество понятий в онтологии составило 578, количество отношений – 15, максимальная вложенность понятий – 12, количество лексических меток на русском языке для каждого элемента онтологии – от 1 до 9.

В онтологию было помещено 1227 экземпляров различных понятий предметной области. Из них 112 экземпляров были выявлены в результате анализа документов, относящихся к выбранной предметной области. Эти экземпляры содержали от 1 до 3 лексических меток. Остальные 1115 экземпляров были автоматически сгенерированы специально созданной программой. Для таких экземпляров использовались лексические метки родительских понятий.

Наполненная экземплярами онтология использовалась для аннотирования 27 рубрик каталога и 160 документов, относящихся к выбранной предметной области. Аннотирование выполнялось в соответствии с методом формирования семантических метаданных. В результате аннотирования объектов портала была выявлена необходимость развития средств визуального представления онтологии и текста документа при составлении семантических метаданных.

Созданные семантические метаданные объектов портала использовались для тестирования остальных функций семантического ядра – поиска, категоризации и формирования рекомендаций.

Указанные функции оценивались тремя тестерами с точки зрения качества с использованием показателей формальной полноты и формальной точности. Показатели названы формальными потому, что они основываются на оценке формальной релевантности, а не на оценке удовлетворенности (рис. 4).

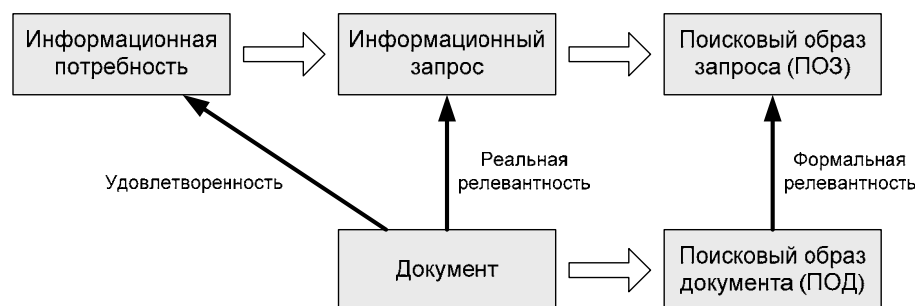


Рис. 4. Схема оценки качества работы ИПС

Нужно отметить, что выбранные показатели нельзя сравнивать с показателями полноты и точности полнотекстовых ИПС в силу того, что формальные показатели не учитывают удовлетворенность пользователей результатами поиска. В данном тестировании удовлетворенность не была оценена потому, что семантические метаданные документов были сгенерированы произвольным образом и не отражали текстового содержания документов. Тем не менее, при достижении высокой степени соответствия семантических метаданных текстовому содержанию документов можно ожидать высоких показателей полноты и точности предложенных методов семантического сравнения.

$$R_F = \frac{a_{md}}{a_{md} + b_{md}} - \text{показатель формальной полноты, где} \quad (8)$$

a_{md} – количество релевантных семантических метаданных, выданных алгоритмом,

b_{md} – количество релевантных семантических метаданных, не выданных алгоритмом.

$$P_F = \frac{a_{md}}{a_{md} + c_{md}} - \text{показатель формальной точности, где} \quad (9)$$

a_{md} – количество релевантных семантических метаданных, выданных алгоритмом,

c_{md} – количество нерелевантных семантических метаданных, выданных алгоритмом.

В процедурах поиска и категоризации используются одинаковые показатели семантической близости элементов онтологии – показатели с учетом наследования, – поэтому для этих процедур были получены одинаковые значения формальной полноты и формальной точности. Средняя формальная полнота составила 1, а средняя формальная точность составила 0,975.

В процедуре формирования рекомендаций используется пороговое значение для рекомендации. От этого значения зависит качество рекомендации. Чем больше пороговое значение, тем выше точность и меньше полнота. И наоборот, чем ниже пороговое значение, тем меньше точность и выше полнота. Было принято решение обеспечить максимальное значение показателя полноты рекомендации. Для этого в результате тестирования было определено пороговое значение для рекомендации, которое составило 0,4469. В этом случае формальная полнота рекомендации равна 1, а формальная точность варьируется от 0,333 до 1 и в среднем составляет 0,5646 (рис. 5).

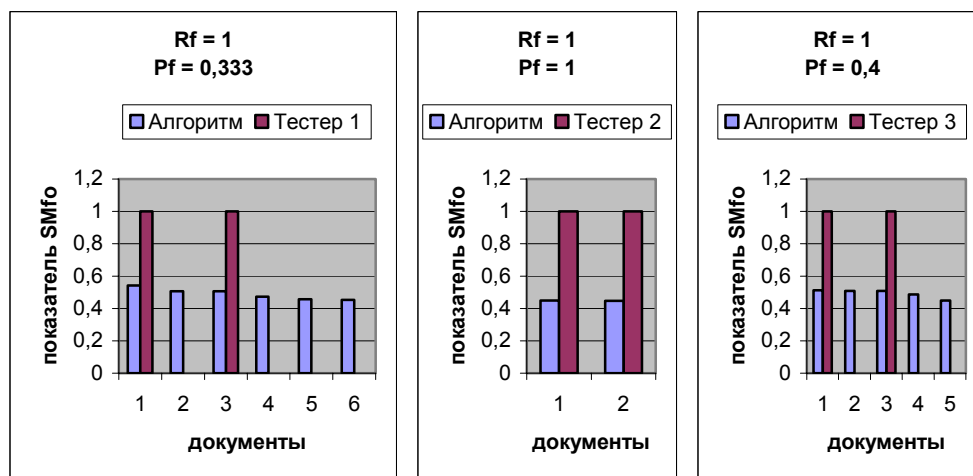


Рис. 5. Сравнение результатов работы тестеров с результатами работы алгоритма для второго тестового документа

Тестирование функций семантического ядра выявило высокие показатели качества их работы.

Разработанное семантическое ядро, реализующее предложенные структуры, методы и алгоритмы, использовалось при создании двух порталов различного уровня.

Портал «Petroleum Engineers Virtual Network» представляет собой программную систему управления явными и неявными знаниями для коллектива специалистов в области разработки нефтяных месторождений. Данный портал был внедрен в Центре профессиональной переподготовки специалистов нефтегазового дела ТПУ (г. Томск). Программная реализация портала была зарегистрирована в Отраслевом фонде алгоритмов и программ.

Корпоративная система управления знаниями разрабатывается в настоящее время для компании «ЭлеСи» (г. Томск). В качестве программно-технической части системы разработан и внедрен семантический портал, функциональность которого основывается на использовании разработанного семантического ядра. В качестве наиболее существенных функциональных подсистем портала реализуются:

- подсистема электронной библиотеки;
- подсистема профилей компетенции ведущих сотрудников компании;
- подсистема поиска;
- подсистема для работы экспертов компании.

ОСНОВНЫЕ РЕЗУЛЬТАТЫ РАБОТЫ

В ходе диссертационного исследования получены следующие основные результаты:

1. Выполнен анализ существующих подходов к разработке семантических порталов. Выявлена доминирующая роль онтологического подхода к созданию семантических порталов. Показано, что с помощью онтологий может решаться широкий круг задач повышения эффективности и качества работы информационных систем.
2. Проведен анализ и обобщение возможных вариантов использования онтологий в информационных системах. Для реализации информационных процессов в портале с учетом семантики объектов предложены варианты использования онтологий.
3. Разработан состав и структура семантического ядра портала. Ядро состоит из сервера онтологий и сервера семантических метаданных. Функциональность семантического ядра портала основывается на логическом формализме представления знаний – дескриптивной логике. В соответствии с указанным

формализмом выбраны языки записи онтологии и семантических метаданных для использования в семантическом ядре портала. Обоснована структура онтологии, обеспечивающая работу семантического ядра портала.

4. Разработан метод формирования семантических метаданных для создания описаний объектов портала. Разработаны методы вычисления семантической близости элементов онтологии и метаданных, формализующие использование понятия сотипности. Указанные методы применены в функциях семантического поиска, категоризации и формирования рекомендаций.
5. Выполнена программная реализация разработанного семантического ядра портала, составившая в общей сложности более 16 тысяч строк кода. Осуществлено тестирование программного кода на сгенерированном множестве семантических метаданных.
6. Разработанные структуры, методы и алгоритмы построения семантических порталов, а также соответствующее программное обеспечение, внедрены в двух организациях (ЗАО «ЭлеСи», Центр профессиональной переподготовки специалистов нефтегазового дела ТПУ) при создании для них семантических порталов различного уровня.

СПИСОК ОПУБЛИКОВАННЫХ РАБОТ ПО ТЕМЕ ДИССЕРТАЦИИ

1. Васильев И. А., Усов М. В. Применение онтологического подхода в информационных системах // Труды X-ой международной научно-практической конференции студентов, аспирантов и молодых ученых «Современные техника и технологии 2004». – Томск: Изд-во ТПУ, 2004. – Том 2. – С. 123-124.
2. Тузовский А. Ф., Васильев И. А. Структура системы управления знаниями // Труды международного симпозиума «Информационные и системные технологии в индустрии, образовании и науке. – Караганда: Издательство КарГТУ, 2003. – С. 286-288.
3. Усов М. В., Васильев И. А. Роль дескриптивной логики в порталах управления знаниями // Труды X-ой международной научно-практической конференции студентов, аспирантов и молодых ученых «Современные техника и технологии 2004». – Томск: Изд-во ТПУ, 2004. – Том 2. – С. 212-213.
4. Васильев И. А. Выбор средства представления знаний для их использования в работе информационного портала организации // Материалы XLII Международной научной студенческой конференции «Студент и научно-технический прогресс»: Информационные технологии. – Новосибирск: Изд-во НГУ, 2004. – С. 56-60.
5. Тузовский А. Ф., Васильев И. А., Усов М. В. Программная реализация основных компонент информационно-программного обеспечения системы управления знаниями // Известия ТПУ. – 2004. – Том 307. – №7. – С. 116-122.
6. Васильев И. А. Оценка семантической близости объектов с использованием дескриптивной логики // Материалы 5-ой научно-практической конференции «Современные средства и системы автоматизации». – Томск: Изд-во ТУСУР, 2004. – С. 160-163.
7. Васильев И. А. Организация коллективной работы пользователей с документами в сети Internet // Материалы XL Международной научной студенческой конференции «Студент и научно-технический прогресс». – Новосибирск: Изд-во НГУ, 2002. – С. 31-32.
8. Васильев И. А., Бубнов Д. В., Козлов С. В. Использование онтологии предметной области для поддержки работы сложных технических систем // Труды IX-ой международной научно-практической конференции студентов, аспирантов и молодых ученых «Современные техника и технологии». – Томск: Изд-во ТПУ, 2003. – Том 2. – С. 144-145.

9. Васильев И. А., Бубнов Д. В., Козлов С. В. Использование сети Интернет для активизации взаимодействия специалистов предметной области // Труды Всероссийской научно-практической конференции «Технологии ИНТЕРНЕТ – на службу обществу» – Саратов: Копипринтер, 2003. – С. 164-166.