На правах рукописи

Иванов Владимир Владимирович

МОДЕЛИ И МЕТОДЫ ИНТЕГРАЦИИ СТРУКТУРИРОВАННЫХ ТЕКСТОВЫХ ОПИСАНИЙ НА ОСНОВЕ ОНТОЛОГИЙ

05.13.11 — Математическое и программное обеспечение вычислительных машин, комплексов и компьютерных сетей

АВТОРЕФЕРАТ
диссертации на соискание ученой степени
кандидата физико-математических наук

Казань — 2009

Работа выполнена на кафедре теоретической кибернетики государственного образовательного учреждения высшего профессионального образования «Казанский государственный университет им. В.И. Ульянова-Ленина»

Научный руководитель: доктор физико-математических наук, профессор Соловьев Валерий Дмитриевич

Официальные оппоненты: доктор технических наук, профессор Гаврилова Татьяна Альбертовна

доктор физико-математических наук, профессор Елизаров Александр Михайлович

Ведущая организация: Научно-исследовательский вычислительный центр МГУ им. М.В. Ломоносова

Защита состоится «4» июня 2009 г. в 16^{00} часов на заседании диссертационного совета Д 212.081.24 при ГОУВПО «Казанский государственный университет им. В.И. Ульянова-Ленина» по адресу: 420008, г. Казань, ул. Кремлевская, д. 35, конференц-зал научной библиотеки им. Н.И. Лобачевского.

С диссертацией можно ознакомиться в научной библиотеке им. Н.И. Лобачевского Казанского государственного университета.

Автореферат разослан «___» апреля 2009 г.

Ученый секретарь
диссертационного совета,
к. ф.-м. н., доцент Еникеев А.И.

Общая характеристика диссертации

Актуальность работы. В диссертации ставятся и решаются задачи, связанные с разработкой математического и программного обеспечения процессов интеграции структурированных текстовых описаний на основе прикладной онтологии. Создание универсальных подходов и методов для интеграции и доступа к описаниям делает возможным применение разработанных методов в слабоформализуемых предметных областях, где необходима обработка смыслового содержимого структурированных текстовых данных.

В теории реляционных баз данных (БД) задачи интеграции поставлены достаточно давно, однако методы семантической интеграции на основе формальных моделей предметных областей (*онтологий*) стали развиваться относительно недавно (5–10 лет назад). Разработаны инструментальные среды для поддержки процесса интеграции, но качество их работы существенно зависит от уровня детализации используемой онтологии и компетентности эксперта, выполняющего интеграцию. Результатом работы автоматизированных систем интеграции является отображение, или т. н. *мэппинг* (от англ. mapping) между структурами разных БД (*схемами*). Предложено множество подходов, использующих для построения отображения как схему БД, так и содержимое. Однако слабо развиты подходы, направленные на спецификацию с помощью онтологий непосредственно содержимого разнородных баз данных с целью интеграции на уровне экстенсионала, т. е. утверждений об объектах предметной области и их свойствах, представленных в БД.

Задачи интеграции структурированных описаний возникают также при представлении информации в глобальной сети Интернет, где очевидны непригодность языка XHTML и недостаточная специализация выразительных средств языка XML для формального описания смыслового содержимого ресурсов веба. С появлением спецификаций языков описания ресурсов (RDF/RDFS) и языка представления знаний (OWL) актуальной проблемой стала реализация идеи семантического веба (Semantic Web, или Web 3.0), которая основана на автоматической обработке смыслового содержимого ресурсов веба по их онтологическому представлению. На данный момент решение этой задачи затруднено из-за слабой развитости методов выражения содержимого существующих ресурсов с помощью веб-онтологий, что объясняется трудоемкостью обработки сплошных текстовых данных при заполнении онтологии фактами. Поставленные в диссертации задачи актуальны при реализации идей семантического веба на основе

множества структурированных описаний, динамически генерируемых из онлайновых БД, т. н. «глубинного веба» (Deep Web), объемы которого во много раз превосходят объемы статического веба.

Потребность в обработке структурированных текстовых описаний возникает, в частности, в области культурного наследия, которая имеет широкий охват как по терминологии, используемой в текстах описаний, так и по разнообразию структур описаний. К данной области обычно относят такие категории организаций, как архивы, библиотеки и музеи, причем для музеев задачи интеграции представляют особую сложность в силу разнообразия и разнотипности описываемых объектов. В музеях России широко внедряются и используются автоматизированные информационные системы (АИС), ориентированные на поддержку учета коллекций и фондов, организацию электронного документооборота, каталогизацию, обработку учетно-хранительской документации. В базах данных музейных АИС содержится большинство электронных описаний предметов музейного фонда России. Несмотря на то, что по оценке Министерства культуры РФ в электронном виде представлены описания более 5 млн. музейных предметов, эта информация используется в основном внутри музеев. Поддержке другой основной функции музеев – обеспечению доступа широкой аудитории к информации по культурному наследию – разработчики отечественных АИС уделяют недостаточное внимание в силу следующих проблем.

1. *Разнородность структур музейных БД* порождается тем, что каждый музей имеет свои особенности, требования и ограничения, а схема БД АИС варьируется от музея к музею.
2. *Разнородность терминологии* порождается тем, что для описания одних и тех же сущностей в разных музеях используются различные системы терминов, в результате чего справочники различных БД существенно отличаются и не могут быть использованы при интеграции.
3. Фактическое *отсутствие единого стандарта* на разделяемую большинством музеев концептуальную модель представления информации о культурном наследии обусловлено наличием нескольких независимых и не связанных друг с другом концептуальных моделей.

В решении этих проблем за рубежом достигнуты определенные результаты: созданы крупные (англоязычные) терминологические ресурсы (например, словари фонда П. Гетти), стандартизирована концептуальная модель

CIDOC CRM, предназначенная для интеграции данных в сфере культурного наследия. Открытым остается вопрос о методологии интеграции разнородных музейных описаний на основе указанных ресурсов, создании модели процесса интеграции, который бы учитывал и структуру, и содержимое музейных описаний. Разработка методов семантической интеграции структурированных текстовых описаний позволит создать единый интерфейс для доступа к описаниям музейного фонда в целом, что весьма востребовано при формировании единого каталога музейных предметов.

Методы и модели интеграции, предложенные в диссертации, использовались для автоматизированного заполнения базы знаний фактами, извлеченными из структурированных текстовых описаний (представлений, построенных над музейными БД) различной структуры. Доступ к базе знаний осуществляется на основе технологий информационного поиска.

Цель и основные задачи. Цель диссертации состоит в разработке математического, программного и лингвистического обеспечения систем семантической интеграции структурированных текстовых описаний. Для достижения цели были поставлены и решены следующие *основные задачи.*

1. Создание прикладной онтологии на основе онтологии верхнего уровня и информационно-поискового тезауруса (ИПТ).
2. Разработка и реализация модели процесса интеграции разнородных структурированных текстовых описаний. Данная задача распадается на две подзадачи:
   – разработку методов спецификации структуры и содержимого описаний с помощью онтологии;
   – разработку методов автоматизированного построения и оценки отображения структурных элементов и текстового содержимого описаний на онтологию.
3. Разработка алгоритма поиска в интегрированном хранилище описаний по запросу на языке, близком к естественному.
4. Проведение экспериментов (на примере области культурного наследия) для оценки качества предлагаемых моделей и методов.

Объект исследования. Структурированные текстовые описания, онтологии верхнего уровня, ИПТ, базы декларативных знаний.

Предмет исследования. Методы семантической интеграции разнородных структурированных текстовых описаний на основе прикладной онтологии.

Методы исследования. При выполнении работы использованы методы, разработанные в области интеграции данных, информационного поиска, машинного обучения и онтологического инжиниринга, описанные в работах отечественных и зарубежных ученых: Д.А. Поспелова, Т.А. Гавриловой, Б.В. Доброва, Г.С. Осипова, В.Ф. Хорошевского, Н.В. Лукашевич, С.Д. Кузнецова, Н. Гуарино, Н. Ной, Т. Грубера, Т. Бернерса-Ли, Д. МакГиннесс, Ф. Баадера, Д. Фенселя и др., а также элементы теории графов и математической логики.

Научная новизна работы. Научной новизной обладают следующие элементы диссертации:

1) подход к формализации связей между ИПТ и онтологией верхнего уровня в виде логических ограничений;
2) алгоритм поиска элементарных соответствий между элементами схем структурированных описаний на основе анализа текстового содержимого элементов и техники латентного семантического анализа;
3) подход к разрешению лексической многозначности и результаты ее экспериментального исследования в структурированных текстовых описаниях.

Практическая значимость. Результаты диссертации могут быть использованы в дальнейших исследованиях в области организации баз данных и знаний, технологий семантического веба, а также при решении практических задач в области интеграции музейных описаний, например, для создания сводных каталогов музейных предметов. Результаты работы использовались в учебном процессе в Казанском государственном университете при чтении курса «Онтологии и тезаурусы» и в Казанском государственном университете культуры и искусств при чтении курсов «Информационные технологии и технические средства в музейном деле» и «Компьютеризация музейных фондов».

Результаты, выносимые на защиту.

1. Подход к созданию прикладной онтологии как концептуальной основы для проектирования базы знаний, основанный на связывании онтологии верхнего уровня и тезауруса с помощью логических ограничений.
2. Модель процесса интеграции разнородных структурированных текстовых описаний для заполнения фактами единой базы знаний.
3. Метод поиска соответствий между элементами структурированного

описания и онтологией, а также подход к разрешению лексической многозначности в базе знаний, которая построена при интеграции разнородных описаний из музейных баз данных.

Апробация результатов работы. Результаты работы докладывались на следующих конференциях:

– Electronic Information, the Visual Arts & Beyond (EVA Moscow) в 2005 – 2007 годах;
– Theory.Engineering.Language (TEL, Казань) в 2006 – 2008 годах;
– Европейской конференции по искусственному интеллекту (ECAI, Рива-дель-Гарда, Италия) в 2006 году;
– конференции Комитета по документации международного совета музеев ICOM (CIDOC, Вена, Австрия) в 2007 году;
– конференции «Знания – Онтологии – Теории» (ЗОНТ, Новосибирск) в 2007 году;
– конференции по когнитивной науке (COGSCI, Москва) в 2008 году;
– совместных семинарах факультета ВМК КГУ и НИИММ им. Н.Г. Чеботарева по перспективным информационным технологиям в 2007 и 2008 годах;
– Казанском научном семинаре «Методы моделирования» в 2007 году;
– итоговой научной конференции КГУ за 2006 – 2008 годы.

Структура диссертации. Диссертация состоит из трех глав, введения и заключения, содержит 145 страниц, 20 рисунков, 24 таблицы. Список литературы содержит 94 источника.

Краткое содержание диссертации

Во **введении** описаны проблемы, рассматриваемые в диссертации, обоснована актуальность исследования, сформулированы цели и задачи работы.

Глава 1 содержит обзор общедоступных ресурсов онтологического характера, на основе которых представляется возможным построить онтологию для интеграции структурированных текстовых описаний. Описаны теоретические и технологические аспекты интеграции разнородных баз данных, приведены примеры АИС, использующих онтологии для доступа к информации в сфере культурного наследия.

В качестве типичных представителей ресурсов онтологического типа

рассматриваются онтологии верхнего уровня и информационно-поисковые тезаурусы. Выбор такого рода ресурсов обоснован двумя факторами: необходимостью формального описания свойств и взаимосвязей объектов предметной области и потребностью использования разнообразной терминологии. Рассмотрены следующие онтологии верхнего уровня: SUMO [1][1], DOLCE [2], CYC [3], CIDOC CRM [4] и тезаурусы: тезаурус по архитектуре и искусству AAT [6], тезаурус по искусству и музейному делу, разработанный в Ленинградском государственном институте культуры им. Н.К. Крупской (СПбГУКИ), а также иконографический тезаурус Ф. Гарнье.

Кроме онтологий верхнего уровня и тезаурусов, в главе рассматриваются существующие форматы описания музейных метаданных, используемые в современных музейных АИС. Описаны следующие отечественные и зарубежные форматы и стандарты: краткое описание (этикетка) музейного предмета, рекомендации Российского этнографического музея по составлению научного паспорта музейного предмета [7], рекомендации британского консорциума MDA (http://www.mda.org.uk/spectrum). Сделаны выводы о возможности использования указанных форматов для автоматической обработки содержимого соответствующих им структурированных описаний.

Выполнен обзор теоретических подходов и технологий интеграции разнородных структурированных данных. В этой области выделяют общие направления на основе федеративных БД, медиаторов и хранилищ данных [9]. В [10] отмечается, что важным аспектом при интеграции данных является наличие глобальной концептуальной схемы. В большинстве случаев задачи интеграции данных сводятся к поиску близких по значению элементов схем путем сравнения структур данных [11, 12], реже — путем сравнения содержимого [13]. Широко используются подходы на основе нейронных сетей [14], машинного обучения [15] и информационного поиска [16]. Результатом сравнения схем данных являются наборы соответствий между элементами схем, на основе которых строится отображение. В [17] рассматриваются два подхода к определению таких отображений: LAV (local-as-view) и GAV (global-as-view), различие между которыми состоит в том, элементы какой из схем (глобальной или локальной) используются как атомы при выражении смысла элементов другой схемы. Для задачи сравнения схем данных разработано множество подходов как специфичных для предметной области [18], так и направленных на использование конкретных языков представления схем [19].

Применяемые методы лингвистической обработки основаны на разнообразных идеях от сравнения n-грамм, оценки расстояния

[1] Список использованной литературы приведен в конце автореферата.

редактирования (расстояния Левенштейна) и созвучности до анализа лексического состава [20]. Такие тезаурусы, как WordNet, используются в качестве базы синонимов при сопоставлении лексических меток элементов схем [8], а также для сравнения значений в текстовом содержимом схем.

В конце главы кратко описаны программные системы, успешно использующие онтологии для доступа к информации по культурному наследию. Среди них – финский портал Finnish Museums on the Semantic Web (http://www.museosuomi.fi/), голландская система MultimediaN (http://e-culture.multimedian.nl/) и проект единой Европейской библиотеки (http://europeana.eu). Проект SCULPTEUR (http://www.sculpteurweb.org/), охватывает 6 европейских музеев с обширными коллекциями цифровых изображений, видеоматериалов с текстовым описанием и метаданными. Система поиска дает возможность пользователю получать доступ к коллекции по комбинации текста, метаданных и концептов онтологии. Рассмотрены также проекты MINERVA, MICHAEL, BRICKS. В этих проектах основной формой представления описаний являются таблицы, содержащие в ячейках текстовые данные – значения атрибутов тех или иных сущностей.

Глава 2 посвящена разработке методов семантической интеграции и доступа к структурированным текстовым описаниям на основе прикладной онтологии. В начале главы описан оригинальный подход к связыванию онтологии верхнего уровня и информационно-поискового тезауруса, организованного по блочно-фасетному принципу. Оба ресурса (и онтология, и тезаурус) представлены на языке OWL DL, основанном на формализме дескриптивной логики $\mathcal{SHOIN}(\mathcal{D})$ – разрешимом фрагменте логики предикатов. Предложен подход к формализации смысла связи между структурными элементами онтологии верхнего уровня и тезауруса. Базовым структурным элементом в онтологии является класс (множество индивидов или экземпляров), а в тезаурусе – понятие, которое также обозначает множество объектов моделируемого мира, однако не имеет явно задаваемых экземпляров, но может иметь парадигматические связи с другими понятиями тезауруса. Выделены две стратегии связывания ИПТ и онтологии:

1) понятия тезауруса становятся подклассами существующих классов онтологии. С каждым понятием тезауруса может быть связано множество его экземпляров. Новый класс-понятие наследует все формальные свойства суперкласса из онтологии;
2) понятия тезауруса внедряются в онтологию как экземпляры особого

мета-класса онтологии. При этом невозможно описывать экземпляры понятий и их структуру, но возможно моделировать иерархии тезауруса, отношение синонимии, используя метасвойства онтологии.

Как более гибкая была выбрана вторая стратегия. Связывание осуществляется с помощью определения набора логических ограничений, накладываемых на множества допустимых значений формальных свойств, заданных в онтологии верхнего уровня. В качестве множества допустимых значений некоторого свойства P выступают группы близких понятий тезауруса, которые обычно представляются как фасеты или дескрипторные блоки. Логические ограничения имеют следующий вид:

$$C(y) \stackrel{\Delta}{=} \forall x.\, P(y, x) \rightarrow DB(x) \qquad \text{(строгая форма ограничения),}$$

либо

$$C(y) \stackrel{\Delta}{=} \exists x.\, P(y, x) \wedge DB(x) \qquad \text{(ослабленная форма ограничения),}$$

где C – унарный предикат (класс C), P – бинарный предикат (свойство P класса C), а DB – унарный предикат (класс, полученный из фасета или дескрипторного блока тезауруса). В общем случае вместо DB может использоваться предикат, истинный на произвольном подмножестве понятий тезауруса. На языке дескриптивной логики ограничения выражаются следующим образом: $\forall P.DB$ и $\exists P.DB$.

Предложенный подход позволяет явно выражать значение класса онтологии верхнего уровня через подмножество понятий тезауруса, допустимых в качестве значений свойства этого класса, что поддерживает независимое ведение ресурсов и отражает фундаментальное разделение между интенсинальной (*т. е. схемой*) и экстенсиональной (*т. е. данными*) компонентами структурированных описаний. Результатом задания конкретных ограничений является прикладная онтология, которая определяет структуру базы знаний и используется для решения задач интеграции и поиска информации.

Дальнейшее изложение во второй главе строится в соответствии со следующими этапами *процесса интеграции* структурированных описаний.

Этап 1. Представление структурированного текстового описания в виде схемы на языке OWL DL.

Этап 2. Поиск множества соответствий между элементами схем. Построение частичного отображения.

Этап 3. Определение полного отображения.

Этап 4. Реализация отображения. Выполнение построенного отображе-

ния и фиксация результата.

Этап 5. Оценка качества результата отображения. При необходимости возможен возврат к этапу 3 для улучшения качества построенного отображения.

Центральным понятием в предлагаемом процессе интеграции является *структурированное описание*, которое моделирует форму представления структурированных текстовых описаний.

Определение 1. Пусть задано множество из n типов данных T_d $(d=1, 2, \ldots, n)$. Тогда R назовем *структурированным описанием* на типах T_d. если оно состоит из двух частей: *интенсионала* (*заголовка* или *схемы* описания) и *экстенсионала* (*содержимого* описания).

1. Интенсионал – множество из m атрибутов вида $A_i:(T_{i_1}, T_{i_2}, \ldots, T_{i_p})$, где A_i – имена атрибутов *структурированного описания* R, а каждый элемент T_{i_j} соответствует некоторому имени типа T_d, $i=1, 2, \ldots, m$, $j=1, 2, \ldots, p$.
2. Экстенсионал – множество, состоящее из строк t, где t является множеством компонентов вида $a_i:(v_{i_1}, v_{i_2}, \ldots, v_{i_p})$, а v_{i_k} – значение одного из типов T_{i_j}, связанных с соответствующим атрибутом A_i, $i=1, 2, \ldots, m$, $j=1, 2, \ldots, p$, $k=1, 2, \ldots, h$.

На этапе 1 происходит приведение интенсионала структурированного описания к виду OWL DL схемы (далее — TBox). Этот этап состоит в определении для каждого атрибута нового класса. Для R создается отдельный класс, имеющий связи со всеми «классами-атрибутами». После создания TBox экстенсионал структурированного описания однозначно переносится в экстенсионал схемы: происходит заполнение ABox.

Наиболее важным в процессе интеграции структурированных описаний является этап поиска соответствий между структурными элементами схем описаний (этап поиска элементарных соответствий).

Определение 2. *Элементарным соответствием* между классами из схем S и T назовем семерку $\langle C_S, Subject_T, Property_T, Object_T, \delta, type, w\rangle$, где C_S – класс из схемы S, $Subject_T, Object_T$ – классы из схемы T, связанные свойством $Property_T$ из схемы T, δ – основа для построения данного соответствия, $type$ – тип связи между классами C_S и $Object_T$, w – вес данного элементарного соответствия.

Каждое элементарное соответствие задает связь между классами C_S и

$Object_T$. Параметр *type* – отношение между C_S и $Object_T$ на домене интерпретации (например, отношение включения или эквивалентности). Параметры $Subject_T$ и $Property_T$ определяют контекст в схеме T, в котором множества экземпляров C_S и $Object_T$ могут быть связаны отношением *type*. Параметр δ указывает, на основе каких компонентов значения построено данное элементарное соответствие между классами C_S и $Object_T$. Назначение параметра δ состоит в том, чтобы моделировать интерпретацию и сравнивать содержимое классов C_S и $Object_T$. Параметр δ, например, может представлять регулярное выражение или набор ключевых слов, содержащихся в текстовых представлениях экземпляров класса C_S и класса $Object_T$. В случае семантической интеграции δ представляет собой список понятий тезауруса, которые описывают экземпляры класса C_S в исходной схеме S и допустимые значения свойства $Property_T$ класса $Subject_T$ в результирующей схеме T. Множество элементарных соответствий определяет отображение между схемами S и T, которое далее называется *частичным отображением*.

Задача построения частичного отображения. Пусть даны исходная схема S и результирующая (глобальная) схема T. Для заданного числового порога $0 \le \theta \le 1$ необходимо построить частичное отображение ϕ, содержащее элементарные соответствия, для каждого из которых выполняются условия:

1) C_S^I и $Object_T^I$ связаны отношением *type* при $I = (\delta, (\cdot)^I)$;
2) $\delta = C_S^I \cap Object_T^I \neq \varnothing$;
3) $w \ge \theta$.

Аналогичным образом определяются элементарные соответствия между бинарными предикатами (свойствами) исходной и результирующей схем и ставится *задача построения частичного отображения бинарных предикатов* из исходной схемы на бинарные предикаты из результирующей.

Для решения поставленной задачи необходимо сравнить интерпретации элементов из схем S и T, т. е определить $I = (\delta, (\cdot)^I)$ для каждого возможного соответствия. Сравнение может выполняться экспертом, понимающим значение, стоящее за символами классов и свойств в схемах, но для автоматизации этого процесса необходимо моделировать интерпретацию *I*. Допущение, лежащее в основе данного подхода к моделированию интерпретации, состоит в том, что совокупность текстовых выражений элементов экстенсионала определяет значение (интенсионал) этого класса. Это значение используется для поиска семантически близких классов в

результирующей схеме T. Для реализации подхода достаточно сделать следующее: *для каждого класса* C_S *из исходной схемы* S *построить список, содержащий те понятия тезауруса, которые встретились в лексическом выражении экстенсионала класса* C_S. Таким образом, интерпретация определяется операционально – через процедуру индексирования текстовых значений с помощью понятий тезауруса. Список понятий определяет интерпретацию класса C_S в терминах информационно-поискового языка тезауруса. Связи между классами и понятиями тезауруса, заданные при создании онтологии, используются для автоматического выделения в схеме T классов $Object_T$, семантически близких классу C_S. Параметры $Subject_T$ и $Property_T$ берутся из соответствующего логического ограничения.

Поскольку поиск элементарных соответствий сводится к оценке близости между текстовыми документами, то далее в рамках этапа 2 рассматриваются альтернативные подходы, используемые в области информационного поиска для вычисления близости между документами, способы выбора множества индексирующих термов (например, на основе ключевых слов), способы назначения весов термов при индексировании и т. д. Особый интерес представляет метод сжатия пространства признаков (термов) с помощью техники скрытого семантического анализа (LSA) [5], которое выполняется с помощью сингулярного разложения матрицы, составленной из векторов, представляющих документы в пространстве термов.

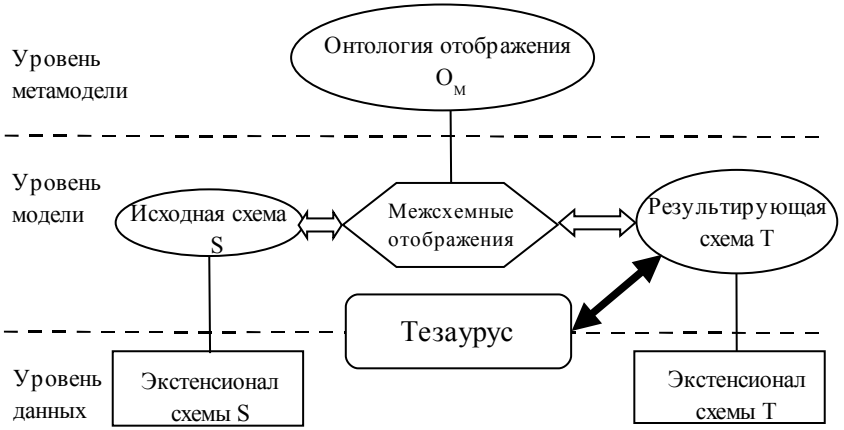

Рис. 1. Логическое представление модели процесса интеграции

В рамках третьего этапа процесса интеграции разработан формат

определения (полного) отображения между схемой исходного источника и результирующей онтологией. Предложена реализация формата определения отображения в виде отдельной метамодели, что позволяет хранить само отображение как независимый набор утверждений и при необходимости использовать их повторно. Метамодель, используемая для описания связей между схемами S и T, представляется на языке дескриптивной логики как *онтология отображения* O_M (рис. 1).

Основным отношением в онтологии O_M является *mapsTo* (*отображаетсяНа*), совокупность значений которого и реализует искомое отображение. Прочие понятия и отношения предназначены для уточнения контекста связываемых элементов разных схем. Задача достраивания частичного отображения до полного основана на критерии связности графа, представляющего запрос к схеме T. Вершинам графа соответствуют классы, дугам – свойства из T. Алгоритм, автоматически достраивающий частичное отображение до полного, должен решать задачу перечисления всех связных подграфов на заданном подмножестве вершин графа и будет иметь экспоненциальную сложность. Для сокращения перебора предлагается использовать следующую эвристику: *ограничивать сверху диаметр графа (т. е. длину максимального кратчайшего пути), моделирующего запрос к схеме T.*

Этап 4 основан на алгоритме, использующем определение полного отображения (O_M) для переноса экземпляров из схемы S в схему T. Данный алгоритм материализует отображение и наполняет базу знаний отдельными фактами, извлеченными из исходного структурированного текстового описания. Одновременно с переносом экземпляров происходит индексирование текстовых значений понятиями тезауруса.

Этап оценки полного отображения основан на следующих общих требованиях. Во-первых, оценка отображения должна учитывать качество результата, который достигается при выполнении отображения. Для оценки результата могут использоваться стандартные подходы на основе критериев точности и полноты. Во-вторых, отображение строится из набора элементарных соответствий, следовательно, оценка качества отображения должна зависеть от входящих в его состав элементарных соответствий. В-третьих, если некоторые из элементарных соответствий не включены в отображение, они должны учитываться при оценке качества, поскольку указывают на то, какая информация теряется при выполнении данного отображения. Обозначим через ξ_i элементарные соответствия, входящие в состав отображения $(i=1,\dots,k)$, а через $\tilde{\xi}_i$ – элементарные соответствия, не включенные в состав отображения $(j=1,\dots,h-k)$. При выполнении

отображения в базу знаний добавляются новые наборы триплетов (троек вида «объект – свойство – значение»), порожденные элементарными соответствиями. Все множество созданных при заполнении базы знаний триплетов можно оценить с точки зрения точности (отношения числа «правильных» триплетов к общему числу сгенерированных триплетов) и полноты (отношения общего числа созданных триплетов к числу значений, реально присутствующих в текстовом выражении экстенсионала исходного класса). Значения критериев точности (P) и полноты (R) комбинируются с помощью формулы F-меры: $F(\xi)=2\mathrm{PR}/(P+R)$. Для оценки качества результата отображения на этапе 5 предложена следующая формула:

$$Q(\Phi)=\alpha(\Phi)(\sum_{i=1}^{k} w_i F(\xi_i))-\beta(\Phi)(\sum_{j=1}^{h-k} \tilde{w}_j F(\tilde{\xi}_j)),$$

где $\alpha(\Phi)$ и $\beta(\Phi)$ – параметры, зависящие от отображения, w_i и $\tilde{w}_j$ – значения веса для элементарных соответствий ξ_i и $\tilde{\xi}_i$.

Для валидации результата отображения может быть использован метод фактографического поиска описаний в базе знаний. Запрос формулируется как набор слов естественного языка и обрабатывается с помощью тезауруса в соответствии с булевской моделью поиска. Каждое понятие тезауруса, извлеченное из текста запроса, сопоставляется с экземпляром онтологии и используется для построения *окрестности* в базе знаний (т. е. связного множества триплетов). При построении окрестности иерархия тезауруса используется естественным образом для расширения запроса. На заключительном шаге алгоритма поиска строится пересечение извлеченных окрестностей (в общем случае может быть использована произвольная логическая формула, включающая основные теоретико-множественные операции, применяемые к окрестностям). Этот же подход лежит в основе построения индекса для ускорения выполнения запросов к базе знаний. Разработан и реализован соответствующий алгоритм поиска по запросу на языке, близком к естественному, учитывающий семантическую разметку. Проведено сравнение алгоритма поиска с одной из классических поисковых машин (ИПС Google) на 8000 описаний музейных предметов. Точность предлагаемого алгоритма поиска на 300 случайных запросах, содержащих понятия тезауруса, увеличивалась в среднем на 11 – 49%.

При индексировании текстовых значений понятиями тезауруса в базе знаний могут возникнуть противоречия, которые порождаются многозначностью лексических единиц тезауруса. Поэтому валидация интегрированной базы знаний опирается также на оценку числа случаев

лексической многозначности (*конфликтов*). Для поиска соответствующих конфликтов в базе знаний используется следующий подход.

Пусть *i* обозначает экземпляр некоторого класса C, P – некоторое свойство класса C, а c_1 и c_2 – экземпляры, представляющие понятия тезауруса. Предикат, описывающий случаи многозначности, определяется следующим образом:

$$Ambig(i, P, c_1, c_2) = P(i, c_1) \wedge P(i, c_2) \wedge Conflict(c_1, c_2),$$

где $Conflict(c_1, c_2)$ принимает истинное значение тогда и только тогда, когда c_1 и c_2 имеют одинаковые текстовые входы в тезаурусе. Для разрешения конфликта необходимо либо отбросить один из $P(i, c_k)$, либо уменьшить область определения $Conflict(c_1, c_2)$.

В заключение главы описаны виды логического вывода, которые возможно реализовать над построенной базой знаний.

1. *Вывод на структуре классов и свойств формальной онтологии.* Возможно конструирование любых правил вывода, поддерживаемых стандартными средствами обработки онтологий, в частности, вывод по транзитивности.

2. *Вывод значений свойств экземпляра по иерархии тезауруса.* Пример правила вывода: $((x, p, y) \wedge (y, BT, z)) \Rightarrow (x, p, z)$, где *y, z* – понятия тезауруса, BT – свойство, представляющее отношение частичного порядка на множестве понятий тезауруса, *x* – экземпляр некоторого класса в базе знаний, *p* – некоторое формальное свойство.

3. *Вывод значений одних свойств объекта по значениям других его свойств.* Пример правила вывода:

$((x, имеетТип, КАРТИНА) \wedge (x, имеетТип, ЛЕС)) \Rightarrow (ЛЕС, изображенНа, x)$.

4. *Вывод новых ассоциативных связей между понятиями тезауруса.*

Пример правила вывода: $((x, p_1, y) \wedge (y, p_2, z)) \Rightarrow (x, связанС, z)$.

В **главе 3** представлены результаты экспериментального исследования алгоритма поиска элементарных соответствий и результаты разрешения лексической многозначности в базе знаний.

Исходными данными при проведении экспериментов являются структурированные описания предметов, извлеченные из трех музейных баз данных: БД ВРМ – Всероссийский реестр музеев, БД ЭМКУ – Этнографический музей Казанского университета, БД РБМ – Рыбинский государственный историко-архитектурный и художественный музей-заповедник.

Приведены количественные характеристики этих источников данных и примеры описаний. Проведены эксперименты с алгоритмом поиска элементарных соответствий, позволяющие судить о качестве работы алгоритма по двум критериям: полноте и точности. График зависимости точности от полноты приведен на рис. 2. Исследованы зависимости критериев точности и полноты от следующих параметров алгоритма.

1. Способ индексирования содержимого (параметр idx) на основе:
   - словоформ (wordform);
   - начальных форм слов (lemma);
   - понятий тезауруса (thesaurus);
2. Способ назначения весов термов (параметр wgt) на основе:
   - признака вхождения терма в документ (binary);
   - числа вхождений терма в документ (count);
   - величины TF*IDF (tf*idf).

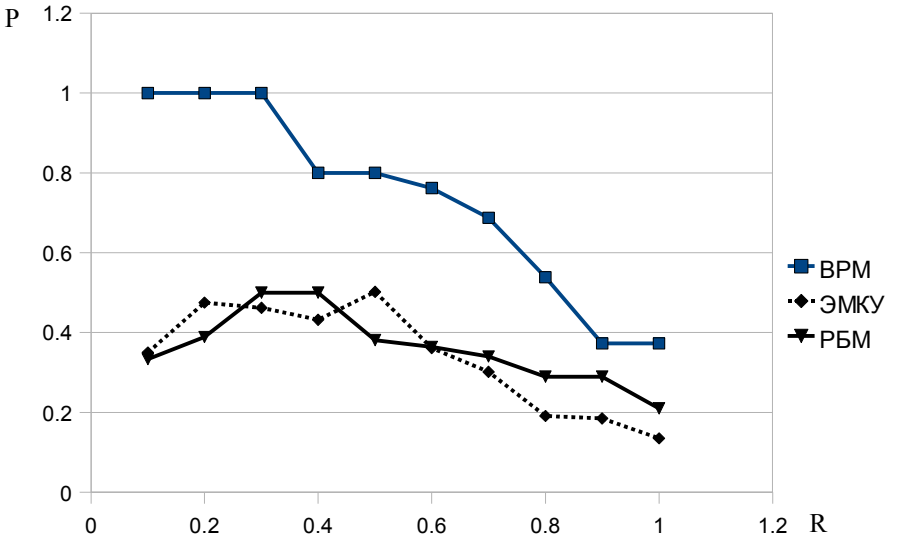

Рис. 2. Соотношение средней точности и полноты при поиске элементарных соответствий для трех БД: ВРМ, ЭМКУ, РБМ

Наилучшие результаты по двум критериям получены для следующих

значений параметров idx={lemma, thes}, wgt={tf*idf}. Для БД ВРМ достигнуты значения P = 80% при R = 56%, а максимальная полнота (100%) достигнута при точности в 60%. Для двух других БД значения критериев были существенно ниже — на уровне 30 – 50% по точности при 60 – 80% по полноте. Такие низкие показатели, однако, не означают, что метод не подходит для автоматизации поиска соответствий. Действительно, если рассмотреть список из всевозможных элементарных соответствий, то точность (на таком списке) для БД РБМ и ЭМКУ будет менее 10%, т. е. алгоритм сокращает число вариантов, которые необходимо рассмотреть эксперту, в 3 – 5 раз. Отмечено, что большая часть элементарных соответствий находится в интервале $0 \le \theta \le 0.05$.

Оценивалось влияние размерности k пространства признаков (термов) на качество работы алгоритма поиска элементарных соответствий. Проведенные эксперименты показали, что использование техники LSA для сжатия пространства термов с помощью сингулярного разложения ведет к увеличению критерия полноты при ухудшении критерия точности. Размерность k влияет на выбор порога θ: при уменьшении размерности до 30 большая часть элементарных соответствий находится в интервале $0 \le \theta \le 0.5$.

Проведено сравнение предлагаемого алгоритма поиска соответствий с известными методами классификации. Для экспериментов были выбраны метод K-ближайших соседей (KNN), основанный на предварительном обучении, и метод кластеризации без предварительного обучения – K-средних (KMeans). Анализ результатов экспериментов показал, что предлагаемый в диссертации метод поиска элементарных соответствий дает лучшие результаты, чем метод KMeans, и по точности (в среднем на 15%), и по полноте (в среднем на 10%), но при этом уступает методу KNN по точности в среднем на 10 – 20%.

Отдельное место в диссертации отводится исследованию лексической многозначности, оказывающей существенное влияние на качество работы предлагаемых методов. Предложено разрешать лексическую многозначность значений в столбце, предварительно определив множество допустимых значений атрибута (домен) как подмножество понятий тезауруса (*нормативный* подход к снятию многозначности).

Эксперименты по снятию многозначности при обработке музейных описаний показали (табл. 1), что учет всех понятий тезауруса AAT при индексировании различных доменов приводит к показателям многозначности на уровне 13 – 30% от общего числа проиндексированных значений, причем большинство случаев многозначности порождается небольшим числом лексических единиц тезауруса из разных фасетов. При применении

нормативного подхода многозначность была в интервале от 0 до 13%. Важно отметить существенное уменьшение полноты покрытия текстовых значений понятиями тезауруса, что наиболее заметно для столбца «Техника».

Для одних столбцов уменьшение числа многозначных единиц в фасете влечет уменьшение числа случаев многозначности при индексировании, для других эта связь нехарактерна, поскольку при описании наименований и типов музейных предметов используются многозначные (внутри фасета) лексические единицы. Поэтому уменьшение числа многозначных единиц с помощью нормативного подхода не приводит к сокращению случаев многозначности при индексировании содержимого столбцов. Для столбцов «Материал» и «Техника», напротив, многозначность порождается *межфасетными* пересечениями, отбрасывая которые, можно существенно сократить число случаев многозначности при незначительном уменьшении полноты индексирования.

Таблица 1

Результаты экспериментов для БД ЭМКУ и БД ВРМ; ALL – индексирование всеми понятиями, NORM – *нормативный* подход

Критерии	Материал		Тип, Название		Техника		БД
	ALL	NORM	ALL	NORM	ALL	NORM	
Полнота, %	98	96	42	38	85	50	ЭМКУ
Многозначность, %	32	1	17	13	31	0	
Количество многозначных единиц	15	2	42	32	26	0	
Полнота, %	97	93	45	43	97	24	ВРМ
Многозначность, %	30	1	13	11	30	0	
Количество многозначных единиц	44	2	53	44	43	0	

В слабоформализованных предметных областях лексическая многозначность порождается метонимами. В частности, в музейной документации термины, обозначающие тип предмета, часто используются для обозначения техники или процесса создания предметов этого типа. Метонимия неявно переносится в тезаурус, а затем и в прикладную онтологию. В этом случае терминологический ресурс, используемый для автоматизированной обработки описаний, должен содержать дополнительно отношение метони-

мии. Добавление этого отношения имеет практическое значение, т.к. опираясь на явное отношение метонимии, можно обосновать корректность некоторых из «ошибочных» элементарных соответствий.

Список публикаций по теме диссертации

Публикации в рецензируемых журналах, рекомендованных ВАК

1. Иванов В.В. Онтологический подход к созданию информационной системы по культурному наследию // Учёные записки Казанского государственного университета. Серия физико-математические науки. – Казань: Казанский государственный университет, 2007. – Т. 149, кн. 2. – С. 73–92.
2. Иванов В.В., Поляков В.Н., Соловьев В.Д. Обзор онтологий верхнего уровня // Вестник Казанского государственного технического университета им. А.Н. Туполева. – 2006. – №3. – С. 50–63 (автором написано 0,7 п. л.).

Прочие публикации

3. Ivanov V. Integration of thesaurus and ontology for the use in the information resource on the culture heritage // Proceedings of First Workshop on Intelligent Technologies for Cultural Heritage Exploitation at the 17th European Conference on Artificial Intelligence. – Trento, 2006. – P. 31–36.
4. Иванов В.В. Использование лингвистических ресурсов для интеграции разнородной музейной документации // Труды Всероссийской конференции с международным участием "Знания–Онтологии–Теории". – Новосибирск: Институт математики им. С.Л. Соболева СО РАН, 2007. – Т. 1. – С. 246–253.
5. Иванов В.В., Соловьев В.Д. Создание и валидация онтологии в области культуры на базе онтологии верхнего уровня и тезауруса // Труды Казанского научного семинара "Методы моделирования". – Казань: Изд-во КГТУ, 2007. – Вып. 3. – С. 135–152 (автором написано 0,8 п. л.).
6. Иванов В.В., Соловьев В.Д. Применение онтологий для разрешения лексической многозначности в структурированных источниках данных // Третья международная конференция по когнитивной науке. – М.: Художественно-издательский центр, 2008. – Т. 2. – С. 577–580 (автором написано 0,2 п. л.).
7. Добров Б.В., Иванов В.В., Лукашевич Н.В., Соловьев В.Д. Онтологии и тезаурусы: Учебно-методическое пособие – Казань: Казанский государственный университет, 2006. – 198 с. (автором написано 6,5 п. л.).
8. Иванов В.В., Соловьев В.Д. Использование онтологий для описания знаний о культурном наследии (обзор работ) // Современный музей как важный ресурс развития города и региона: Материалы международной научно-практической конференции. – Казань: РИЦ «Школа», 2005. – С. 42–46 (автором написано 0,2 п. л.)
9. Добров Б.В., Иванов В.В., Лукашевич Н.В., Соловьев В.Д. Формирование линг-

вистического обеспечения информационной системы по культурному наследию // Сборник трудов конференции «Научный сервис в сети Интернет: технологии параллельного программирования», Новороссийск. – 2006. – С. 257–259 (автором написано 0,05 п. л.)

10. Добров Б.В., Лукашевич Н.В., Иванов В.В. Лингвистическое обеспечение информационной системы по культурному наследию // III Международные Бодуэновские чтения: И.А. Бодуэн де Куртенэ и современные проблемы теоретического и прикладного языкознания: труды и материалы: в 2 т. – Казань: Казанский государственный университет, 2006. – Т. 2. – С. 169–171. (автором написано 0,05 п. л.).
11. Добров Б.В., Иванов В.В., Лукашевич Н.В., Соловьев В.Д. Онтологии и тезаурусы: модели, инструменты, приложения: учебное пособие. – М.: Интернет-Университет Информационных Технологий; БИНОМ. Лаборатория знаний, 2009. – 173 с.: ил. (автором написано 5,5 п. л.).
12. Ivanov V. Integrating heterogeneous museum descriptions using linguistic resources // Proceedings of CIDOC–2007 Conference. – Vienna, Austria, 2007 [Электронный ресурс]. – Режим доступа: http://cidoc.mediahost.org/content/archive/cidoc2007/papers/Ivanov_CIDOC_2007_full_text.pdf, свободный.
13. Иванов В.В. Подход к интеграции разнородных описаний музейных предметов // Сборник тезисов конференции АДИТ–2007, Саратов [Электронный ресурс]. – Режим доступа: http://adit.association.museum/rus/conference/adit2007/papers/paper.asp?nomer=57, свободный.
14. Иванов В.В., Соловьев В.Д. Информационная система "Культурное наследие России" // Труды Международной конференции "EVA–2005 Москва", 2005 [Электронный ресурс]. – Режим доступа: http://conf.cpic.ru/upload/eva2005/reports/doklad_686.doc, свободный.
15. Иванов В.В. Разработка лингвистического ресурса для информатизации музеев // Труды Международной конференции "EVA–2006 Москва" [Электронный ресурс]. – Режим доступа: http://conf.cpic.ru/eva2006/rus/reports/report_839.html, свободный.

Литература

1. Pease A., Niles I. Toward a Standard Upper Ontology // Formal Ontology in Information Systems. Proceedings of the 2nd International Conference (FOIS-2001) / Ed. by C. Welty, B. Smith. – New York: ACM Press, 2001. – P. 2–9.
2. Masolo C., Borgo S., Gangemi A., Guarino N. et al. WonderWeb Deliverable D18 Ontology Library (final). IST Project 2001-33052 WonderWeb: Ontology Infrastructure for the Semantic Web [Электронный ресурс]. – Режим доступа: www.loa-cnr.it/Papers/D18.pdf, свободный.
3. Lenat D.B., Guha R.V. Building Large Knowledge-Based Systems: Representation and Inference in the Cyc Project. – Addison-Wesley, 1990. – 372 p.
4. Crofts N., Doerr M., Gill T., Stead S. Definition of the CIDOC Conceptual Reference Model [Электронный ресурс]. – Режим доступа: http://cidoc.ics.forth.gr/docs/

cidoc_crm_version_4.0.pdf, свободный.
5. Dumais S.T. et al. Using latent semantic analysis to improve access to textual information // Proc. Conf. on Human Factors in Computing Systems, 1988. – P. 281–286.
6. Petersen T., Barnett P. Art & Architecture Thesaurus: Guide to Indexing and Cataloging With the Art & Architecture Thesaurus. – Oxford: Oxford University Press, 1994.
7. Атрибуция музейного памятника: справочник / Под ред. И.В. Дубова. – СПб.: Лань, 1999. – 346 с.
8. Embley D.W., Jackmann D., Xu L. Multifaceted Exploitation of Metadata for Attribute Match Discovery in Information Integration // Proceedings of Intl. Workshop on Information Integration on the Web (WIIW). – 2001. – P. 110–117.
9. Fundamentals of Data Warehousing / Ed. by M. Jarke, M. Lenzerini, Y. Vassiliou, P. Vassiliadis. – Springer-Verlag, 1999.
10. Wache H., Vogele T., Visser U., Stuckenschmidt H. et al. Ontology-Based Integration of Information – A Survey of Existing Approaches // Proceedings of the IJCAI–2001 Workshop: Ontologies and Information Sharing. – Seattle, WA, 2001. – P. 108–117.
12. Do H.H., Rahm E. COMA – A System for Flexible Combination of Schema Matching Approach // Proceedings of Intl. Conference on Very Large Databases (VLDB). – 2002. – P. 610–621.
13. Doan A.H., Madhavan J., Domingos P., Halevy A. Learning to Map between Ontologies on the Semantic Web // Proceedings of Intl. Conference World Wide Web (WWW). – 2002. – P. 662–673.
14. Li W.S., Clifton C., Liu S.Y. Database Integration Using Neural Networks: Implementation and Experiences // Knowledge and Information Systems. – 2000. – V. 2. – №1. – P. 73–96.
15. Berlin J., Motro A. Database Schema Matching Using Machine Learning with Feature Selection // Proceedings of Intl. Conference Advanced Information Systems Engineering (CaiSE). – 2002. – P. 452–466.
16. Cohen W. Integration of Heterogeneous Databases Without Common Domains Using Queries Based on Textual Similarity // Proceedings of ACM SIGMOD Intl. Conference Management of Data. – 1998. – P. 201–212.
17. Baader F., McGuinness D., Nardi D., Patel-Schneider P. The Description Logic Handbook: Theory, implementation and applications. – Cambridge: Cambridge University Press, 2003. – 574 p.
18. Bergamaschi S., Castano S., Vincini M., Beneventano D. Semantic Integration of Heterogeneous Information Sources // Data and Knowledge Engineering. – 2001. – №36(3). – P. 215–249.
19. Miller R.J. et al. The CLIO Project – Managing Heterogeneity // ACM SIGMOD Record. – 2001. – №30(1). – P. 78–83.
20. Xu L., Embley D. Discovering Direct and Indirect Matches for Schema Elements // Proceedings of Intl. Conference on Database Systems for Advanced Applications (DASFAA). – 2003. – P. 39–46.