OSTIS-2012
(Open Semantic Technologies for Intelligent Systems)

УДК 004.82:004.55

WEB-ОРИЕНТИРОВАННАЯ РЕАЛИЗАЦИЯ СЕМАНТИЧЕСКИХ МОДЕЛЕЙ ИНТЕЛЛЕКТУАЛЬНЫХ СИСТЕМ

Колб Д.Г.

Белорусский государственный университет информатики и радиоэлектроники,
г. Минск, Республика Беларусь
kolb@bsuir.by

Рассмотрены принципы и подходы к разработке семантических web-сайтов с использованием средств традиционных технологий построения web-сайтов. В основу предлагаемых подходов положены понятие sc-модели web-сайта и способ псевдоестественного представления таких моделей.
Ключевые слова: семантическая сеть, семантическая модель web-сайта, SC-код, интеллектуальная система.

Введение

Эффективная интерпретация различных моделей интеллектуальных систем является одной из ключевых проблем разработки интеллектуальных систем. Наиболее остро в рамках данного направления стоят проблемы разработки унифицированных подходов и инструментальных средств интерпретации моделей интеллектуальных систем на различные web-платформы.

Базовым структурным элементом при организации любой информационной системы (ИС) на web-платформе, как правило, является web-страница. Web-страница содержит информацию о некотором фрагменте предметной области, которой посвящен web-сайт или web-портал. При такой организации можно выделить уровни, которые содержит любая ИС: ***уровень внешнего представления***, ***уровень языков разметки***, ***уровень серверных скриптов***, ***уровень хранилищ данных.*** Рассматривая в таком ракурсе ИС на базе web-платформы, мы приходим к заключению, что степень интеллектуальности ИС будет зависеть от степени внедрения на каждом из выделенных уровней ИС семантических технологий. Обозначим направления работ по интеллектуализации ИС на базе web-платформы в соответствии с выделенными выше уровнями:

- семантическая структуризация информации на уровне внешнего представления, которая необходима для обеспечения семантически-однозначной и понятной для читателя гипертекста навигации по web-пространству сайта;
- семантическая структуризация информации на уровне языков разметки, которая необходима для обеспечения поисковых машин знаниями, необходимыми для поиска информации на множестве web-ресурсов web-пространства, и которая обеспечивает возможность ввода в интеллектуальную систему на базе web-платформы новых знаний;
- разработка на уровне серверных скриптов операций семантического поиска и навигации по информационному web-пространству сайта для обеспечения пользователей средствами поиска более эффективными, чем обычная навигация, и для обеспечения пользователей средствами решения задач, которые традиционно относят к задачам искусственного интеллекта;
- использование на уровне хранилищ данных хранилищ знаний вместо баз данных, для того, чтобы обеспечить представления информации о предметной области, к которой относится web-сайт, с помощью моделей представления знаний, которые используются в искусственном интеллекте. Такая организация хранилищ знаний позволит эффективно реализовать семантический поиск, а также обеспечит платформу для решения различных задач в рамках предметной области для средств уровня серверных скриптов.

Для обеспечения возможности семантической структуризации в рамках предлагаемого подхода будем рассматривать сущности web-страницы как элементы базы знаний (БЗ). Такой подход позволяет трактовать ИС на базе web-платформы как специализированную интеллектуальную систему, решающую задачу организации диалога человека и предметной интеллектуальной системы, и обеспечивающую решение основных задач предметной интеллектуальной системы. Рассмотрим один из возможных подходов к

решению указанных проблем, который используется в рамках проекта OSTIS.

В качестве формальной основы предлагаемого подхода будем использовать семантические сети с базовой теоретико-множественной интерпретацией. Основным способом кодирования информации для таких сетей является SC-код (Semantic Code) [OSTISa, 2011]. Интеллектуальные системы, построенные с использованием SC-кода, будем называть sc-системами. В основе любой sc-системы лежит понятие некоторой прикладной sc-модели – модели предметной области, закодированной с помощью SC-кода. Таким образом, задача интерпретация некоторой модели интеллектуальной системы на некоторую прикладную (программную или аппаратную) платформу в рамках проекта OSTIS сводится к интерпретации sc-модели этой интеллектуальной системы на данную платформу.

Разработка интеллектуальной системы, при использовании предложенного подхода будет проходить в следующем порядке:

- разработка формального описания предметной области, по которой разрабатывается интеллектуальная система (sc-модели);
- выделение сущностей предметной области, информация о которых будет размещаться на web-страницах;
- разработка операций обработки sc-модели;
- представление информации о выделенных сущностях в формальном виде и погружение этой информации в некоторое хранилище знаний на базе web-платформы;
- тестирование и отладка sc-системы.

Отметим, что приведенные этапы разработки предусматривают реализацию sc-систем с различной функциональностью, что, как правило, обуславливается спецификой конкретной интеллектуальной системы и определяет, один из возможных способов интерпретации этой sc-системы на web-платформу. Например, может быть несколько различных интерпретаций sc-системы, если она принадлежит к классу информационно-справочных систем (к этому классу относится большинство традиционных сайтов). Приведем некоторые из интерпретаций:

- интерпретация sc-системы на базе уровня внешнего представления, которая предполагает ручную семантическую навигацию по web-пространству;
- интерпретация sc-системы на базе уровня внешнего представления и уровня языков разметки, которая предполагает как ручную семантическую навигацию по web-пространству, так навигацию с помощью различных средств семантического поиска;
- интерпретация sc-системы на базе уровня внешнего представления, уровня языков разметки и уровня серверных скриптов, которая предполагает использование функций, доступных на двух предыдущих уровнях и возможность решения задач в рамках предметной области, по которой разработана интеллектуальная система.

Реализация предлагаемого подхода на практике предполагает решение следующего перечня задач:

- разработка языковых средств представления sc-моделей интеллектуальных систем в рамках некоторой прикладной web-платформы, учитывающих особенности представления мультимедиа данных;
- разработка программных средств, поддерживающих реализацию операций обработки sc-моделей;
- разработка программных средств, обеспечивающих эффективное хранение sc-моделей в рамках некоторого высокопроизводительного хранилища данных.
- разработки средств управления жизненным циклом sc-системы, для эффективной поддержки sc-системы на всех этапах её функционирования.

Рассмотрим решение этих задач подробнее.

1. Языковые средства представления sc-моделей интеллектуальных систем

1.1. Состояние работ в области языковых средств представления моделей интеллектуальных систем.

Уровень внешнего представления – это часть web-ресурса, которая видима пользователю. В настоящее время существует масса подходов к структуризации информации на этом уровне web-ресурса. Наиболее известными из них являются:

- структуризация в виде обычного текста;
- структуризация в виде линейного спискового представления;
- структуризация в виде иерархического спискового представления;
- структуризация в виде табличного представления;
- структуризация за счет выделения элементов различной значимости шрифтом, фоном или оформлением;
- структуризация на основе индуктивного подхода к организации web-ресурса [Microsoft, 2001].

Работ, связанных с семантической структуризацией внешнего представления web-ресурсов, в научной литературе практически нет. Однако, существует масса работ по смежной с данной тематикой, когнетике, в которых затрагиваются проблемы представления любых форм информации в рамках пользовательского интерфейса [Раскин, 2004].

Уровень языков разметки – это, в настоящий момент, наиболее проработанный уровень с научной и практической стороны в смысле структуризации и представления данных в рамках web-ресурса. Основные направления работ здесь связаны, с семантической структуризацией текстового и мультимедиа наполнения web-сайтов.

Первыми результатами в данной области можно назвать реализации для сети Internet формализованной модели гипертекста, основу которой составлял язык HTML [W3C, 2011]. Новой ветвью развития языков представления знаний для сети Internet стало направление Semantic web [W3C, 2011]. Однако, до появления этого направления уже существовали работы, которые были направлены на оптимизацию текущего представления знаний в виде HTML для сети Internet [Гаврилова, 2000].

Semantic web предполагает представление знаний в виде семантической сети с помощью онтологий. Основу технологий предлагаемых Semantic Web составляет семейство стандартов на языки описания, включающее XML, XML Schema, RDF, RDF Schema, OWL, OWL2 [Хорошевский, 2008]. Первые результаты работ в рамках направления Semantic web были использованы разработчиками пользовательских интерфейсов web-ресурсов для семантического размещения типовых сущностей пользовательского интерфейса или его структуры. Основу таких подходов составлял язык XML. В настоящее время разработкой языков описания пользовательского интерфейса и технологиями, поддерживающими такие языки, занимаются такие ведущие разработчики инструментальных средств для разработки программного обеспечения, как Microsoft (XAML, MRML), Adobe (MXML, OpenLaszlo), Oracle (CookSwing, SwiXML, SwixNG, Thinlet, Ultrid, Vexi, XALXAL, XSWT, ZUML), Mozilla (XUL).

Однако только семантическая разметка структуры страниц web-ресурса не давала возможность реализовать эффективные способы поиска информации в рамках web-ресурса. Одним из наиболее доступных по реализации подходов для решения данной проблемы стал подход, основанный на использовании микроформатов (microformats, μF или uF) – способа семантически размечать сведения о разнообразных сущностях на web-страницах, используя стандартные элементы языка HTML (или XHTML) [Вики о микроформатах, 2011].

В настоящее время микроформаты используются как для унифицированного представления однотипных сущностей пользовательского интерфейса web-приложения, так и для настройки и адаптации пользовательского интерфейса web-браузеров к пользователю [Веб-фрагмент, 2011].

Несмотря на простоту использования микроформатного подхода, он не позволял решить все задачи, которые ставило перед собой направление Semantic Web. Поэтому одновременно с микроформатным подходом развивались подходы, в основе которых лежит использование RDF. Ярким примером реализации подхода на базе RDF является проект FOAF("Friend of a Friend") [FOAF, 2010], который позволяет описывать отношение знакомства с помощью RDF.

С ростом количества web-ресурсов поддерживающих стандарты Semantic web появилась необходимость унифицированного представления знаний для таких web-ресурсов. Решением такой проблемы является использование системы метаданных для представления знаний в web-ресурсах. В настоящее время существует несколько десятков проектов, связанных с разработкой систем метаданных. Одним из наиболее популярных проектов, направленных на решения проблемы унификации представления знаний в виде семантических сетей, стал проект «Дублинское ядро»[DCMI, 2011]. Целью проекта стала разработка стандартов метаданных, которые были бы независимы от платформ и подходили бы для широкого спектра задач. Основными результатами проекта являются словари метаданных общего назначения, стандартизирующие описание ресурсов с помощью различных RDF-форматов.

Несмотря на наличие серьёзных подходов к структуризации на уровне внешнего представления и уровне разметки многие проблемы остались до сих пор не решенными, а именно:

- отсутствуют средства семантической структуризации внешнего представления web-ресурсов;
- отсутствуют единые стандарты для унификации представления знаний (унифицированных языковых средств разметки). Помимо проекта «Дублинское ядро» в сети Internet широко используются ряд других систем метаданных, таких как GILS, MARC, ONIX, LOM, UDDI;
- отсутствуют унифицированные подходы к представлению мультимедиа данных;
- отсутствуют единые подходы к проектированию интеллектуальных систем, в основе которых лежат семантические сети для сети Internet.

1.2. Языковые средства представления sc-моделей интеллектуальных систем на уровне внешнего представления и уровне языков разметки

Основу уровня внешнего представления в предлагаемом подходе составляют ***семантически структурированные гипертексты*** – гипертексты, информация в которых будет отображаться помощью SCn-кода (способа псевдоестественного кодирования семантических сетей, представленных в SC-коде, Semantic Code natural). SCn-код задается множеством всех sc.n-статей, каждая из которых описывает семантическую окрестность некоторого понятия предметной области. Каждая статья в свою очередь состоит из идентификатора sc-элемента (объекта предметной области, закодированного с помощью SC-кода), описываемого в этой sc.n-статье, и, возможно, одного или нескольких последующих sc.n-полей. При описании sc-элемента в sc.n-статье sc.n-поля описывают как, какими ролями и связками каких отношений, связан описываемый sc-элемент с другими sc-элементами.

Ряд sc.n-полей может содержать мультимедиа или тексты логических утверждений. Мультимедиа может включать любые информационные конструкции, обозначаемые как внешние по отношению к SCn-коду, в том числе и sc.n-тексты.

Основу уровня языков разметки является SCnML (SCn Markup Language) – модель языковых средств разметки семантически структурированных гипертекстов. SCnML – определяет общие правила разметки текстов языка SCn. SCnML содержит следующие классы тегов для разметки sc.n-статей:

- тег описываемого объекта;
- тег связки;
  - •• тег однокомпонентной связки;
    - ••• тег однокомпонентной связки с тегом ролевого отношения;
    - ••• тег однокомпонентной связки без тега ролевого отношения;
  - •• тег многокомпонентной связки;
    - ••• тег многокомпонентной связки с вложением;
    - ••• тег многокомпонентной связки без вложения;
  - •• тег ролевого отношения;
- тег компонента связки;
  - •• тег компонента связки без вложенного ролевого отношения;
  - •• тег компонента связки с вложенным ролевым отношением;
- тег scnml-запроса.

Любая прикладная реализации SCnML должна содержать указанные классы тегов. На множестве scnml-тегов заданы следующие типы отношений: “быть корневым тегом текста, соответствующего sc.n-статье”, “родительский тег - дочерний тег”. Приведем свойства SCnML-текстов:

- каждый SCnML-тег связан с другим SCnML-тегом понятием уровня, уровень позволяет задать отношение между родительским и дочерним тегом и определяет, в рамках какого контекста идет описание стоящее ниже по уровню (для какого понятия определены связки и к какой связке относятся компоненты связок);
- на одной web-странице могут располагаться scnml-тексты, соответствующим нескольким sc.n-статьям;
- scnml-текст, соответствующий одной sc.n-статье может входить в состав scnml-текста, соответствующего другой sc.n-статье;
- теги в рамках scnml-текста записываются по следующим правилам:
  - •• первым следует корневой тег, тег описываемой сущности;
  - •• за корневым тегом может следовать только тег связки;
  - •• тегу компонента связки всегда предшествует тег связки.

Подход к разметке web-страниц на основе SCnML позволяет обеспечить независимость семантической модели предметной области от языка разметки при верстке страниц с семантически-структурированным гипертекстом. Таким образом, обеспечивается решение двух указанных выше задач семантической структуризации на уровне внешнего представления и уровне языков разметки.

1.3. Языковые средства представления мультимедиа данных в sc-моделях интеллектуальных систем

О том, что разработка средств семантической аннотации мультимедиа данных – это востребованная научная и практическая задача, свидетельствуют работы, которые ведутся в рамках проектов Semantic Web. Группой W3C проведен анализ существующих технологий для применения мультимедиа и выделены следующие классы мультимедиа данных: неподвижные изображения, видео, аудио, текст, данные общего назначения. Также выделены основные классы для категоризации использования мультимедиа по следующим направлениям: по рабочим потокам (для публикации и издания и т. д.), по областям применения (развлечения, новости, спорт и т.д.), для промышленного использования (для вещания, для музыки, для издательства и т.д.).

В результате исследования рассмотрены мультимедиа данные, для которых в настоящие время описаны спецификации. Рассмотренные спецификации включают мультимедиа данные, представляемые как с помощью не связанных с XML-платформой средств, так и с помощью средств XML-платформы, и спецификации мультимедиа данных с использованием онтологий. Основной проблемой работ по спецификации мультимедиа данных является наличие узкоспециализированных направлений по разработке мультимедиа спецификаций, которые не связаны друг с другом, о чем можно судить на основании [W3C, 2007].

Работы, связанные с унифицированной спецификацией мультимедиа данных, только начинают развиваться. В рамках данной статьи предлагается один из подходов к универсальной спецификации мультимедиа данных. Основной целью предлагаемого подхода является не разработка неких универсальных средств спецификации мультимедиа данных, а предложение средств, которые позволят, унифицировано, в едином стиле описывать мультимедиа данные, используя уже существующие специализированные системы метаданных.

В качестве языковых средств унифицированной спецификации произвольных мультимедиа данных предлагается использовать язык гипермедийных структур (ЯГС) [Колб, 2009]. ЯГС позволяет специфицировать любой мультимедиа ресурс с учетом семантики предметной области и связей мультимедиа ресурса с понятиями предметной области. ЯГС является sc-языком (то есть языком, построенным на базе SC-кода), ориентированным на представление структурных моделей предметных

областей, в которых объектами исследования являются отображаемые пользователям файлы интеллектуальной системы, а предметом исследования – семейство понятий, обеспечивающих спецификацию этих файлов и описание различных синтаксических и семантических связей между ними.

Уточним некоторые базовые понятия, которые будут использованы в дальнейшем: понятие sc-файла, понятие формата, понятие способ представления формата, способ визуализации формата и области использования формата.

Основным понятием, с которым работает ЯГС, является понятие файла sc-системы – sc-элемента, который является инородным по отношению к SC-коду объектом произвольной структуры. Будем называть файл sc-системы (или файл, который интегрирован в sc-систему) ***sc-файлом.*** В рамках данной работы мы рассматриваем, только те типы sc-файлов, которые являются мультимедийной информационной конструкцией. Такие информационные конструкции в SC-коде обозначаются с помощью sc-ссылок или sc-узлов (в случае, если речь идет о файлах, внешних прикладных программах и внешних программных системах, которые в текущий момент до конца не интегрированы в состав sc-системы – т.е. формально отсутствует сама ссылка) [OSTISa, 2011].

Под ***форматом*** будем понимать спецификацию (однозначное описание) структуры данных, записанных в sc-файле. Формат определяет ***способ хранения*** информационной конструкции (например: растровый, векторный) и ***форму хранения*** информационной конструкции (например: используемый алгоритм сжатия).

Под ***способом представления*** (кодирования) формата будем понимать некоторые языковые средства, позволяющие записать спецификацию структур данных, описываемых некоторым форматом.

Под ***способом визуализации*** формата будем понимать спецификацию информационной конструкции, позволяющую наглядно (с наибольшим когнитивным эффектом) отобразить данную информационную конструкцию.

Под ***областью использования*** формата будем понимать некоторую область информационных технологий в рамках, которой данный формат нашел применение. Как правило, область использования определяет набор требований к формату (качество изображения, наличие или отсутствие алгоритма сжатия, поддержка нескольких аудио-потоков и т.д.).

С точки зрения семантики понятие формата – это небинарное отношение, которое устанавливает связь sc-файла с его способом хранения, формой хранения, способом представления и множеством областей использования формата. Это отношение, позволяющее представить знания о том, как обрабатывать sc-файл, находящий в sc-системе. Каждая связка такого отношения связана отношением ***способ визуализации**** с некоторой информационной конструкцией, которая определяет способ отображения данного формата с помощью средств визуализации и просмотра, которые присутствуют в sc-системе. То есть понятие формата обеспечивает даталогическую (или синтаксическую) спецификацию мультимедиа ресурса. А использование данных отношений для всех форматов позволяет обеспечить унифицированную семантическую спецификацию для всех возможных представлений мультимедиа данных.

Заметим, что sc-файл вовсе не обязан явно “хранить” некоторую информационную конструкцию в качестве своего содержимого. Эта информационная конструкция может быть неизвестна (не сформирована) и разбита на фрагменты, каждый из которых представлен явно, и, следовательно, нет никакой необходимости явно представлять и хранить всю исходную информационную конструкцию. Данное свойство позволяет говорить о возможности унификации различных подходов для анализа мультимедиа данных [Колб, 2011].

Для инфологической (или семантической) спецификации мультимедиа ресурса, используются понятия и отношения, которые находятся в предметной базе знаний sc-системы. Основу ключевых элементов для этого набора отношений задают ключевые узлы языка гипермедийных структур. Основу языка задают классы различных мультимедийных объектов, приведем основные из них:

- Неподвижное изображение
- Абстрактный текст
  - Линейная конфигурация графем
    - Терм
      - Символьное представление числа
      - Текст
  - Нелинейная конфигурация графем
    - sc.g-конструкция
    - sc.s-конструкция
- Аудиоинформация
- Динамическая графика
  - Видеоинформация
  - Анимация
- Контейнер (составной объект, например архив, который может содержать различные мультимедиа-объекты)
- Элемент программного обеспечения

Каждый объект из приведенного набора классов обладает набором уникальных для класса объекта свойств. Например, с объектами класса ***“Неподвижное изображение”*** может быть связана

информация об авторе, дате создания объекта, а также информация о том с помощью какого формата представлен в sc-файле объект. С объектами класса ***"Текст"*** может быть связана информация о кодировке текста, формате представления текстовой информации, авторе, дате создания, дате модификации и т.д.

Приведем некоторые базовые отношения, которые задаются на множестве мультимедийных объектов: ***презентация****, ***иллюстрация****, ***автор****, ***комментарий****, ***пояснение****, ***биография****, ***карта****, ***рецензент****, ***дата создания****, ***дата модификации**** и др.

2. Интерпретация семантической модели web-сайта на различные платформы

2.1. Состояние работ в области языковых средств обработки и эффективного представления в хранилищах данных моделей интеллектуальных систем

Уровень серверных скриптов, предоставляет разработчикам web-ресурсов широкий выбор инструментов, позволяющих проводить различные операции над разметкой и реализовать требующийся от web-ресурса набор функций. Для современных web-сайтов это инструмент, позволяющий динамически изменять разметку и вносить изменения на внешнем представлении, – другими словами – это основной способ реструктуризации на уровне разметки и внешнего представления. Кроме этого этот уровень на сегодняшний день является звеном, обеспечивающим связь уровня разметки и внешнего представления с уровнем хранилищ данных.

Научный и практический потенциал, достигнутый на этом уровне, позволяет говорить о широких возможностях для реализации алгоритмов искусственного интеллекта. Кроме традиционных технологий, используемых в программировании, в ряде языков, используемых в web-разработках, реализована поддержка следующих парадигм программирования:

- функциональной (PHP, Python, Ruby, Perl, Erlang);
- логической (Mercury [Mercury 2011] через API для Java, C#, Erlang);
- декларативной (PHP, Python, Ruby, Perl);
- метапрограммирования на основе интроспекции и интерпретации произвольного кода (Python, Ruby, PHP и ряд других языков).

Уровень хранилищ данных. Всплеск работ по хранилищам данных и знаний, которые ориентированы на отличное от реляционной модели представление произошел в 2009 году с появлением целого ряда практических результатов в рамках направления NoSQL (Not only SQL) [NoSQL, 2011]. Появление работ было связано с тем, что современные реляционные СУБД не могут эффективно справляться с нагрузками, возникающими при обработке огромных объёмов данных в виду наличия ряда проблем в трёх областях:

- горизонтальное масштабирование при больших объемах данных; например, как в случае социальной сети Digg (3 терабайта для зеленых значков, отображаемых, если ваш друг сделал "dugg" на статье) или в случае социальной сети Facebook (50 терабайт для поиска по входящим сообщениям), или в случае системы Internet-аукционов eBay (2 петабайта в целом);
- производительность каждого отдельного сервера;
- не гибкий дизайн логической структуры [Новожилов, 2011].

Как средство, эффективно решающее проблемы реляционных СУБД, появились различные сетевые СУБД, объектно-ориентированные СУБД, а также СУБД "ключ значение".

Отметим основные недостатки двух нижних уровней:

- высокая степень зависимости существующих технологий разработки web-приложений от платформ, на которых они реализованы, что приводит к существенным изменениям этих технологий при переходе на новые платформы;
- стремление современных подходов к разработке web-ресурсов обеспечить высокий уровень гибкости приложений, без уделения достаточного внимания принципам унификации и интеграции web-ресурсов, построенных на базе разных подходов.

2.2. Иерархическая модель семантического web-сайта

Формальной основной предлагаемого подхода выбран SC-код, поэтому отправной точкой при построении интеллектуальной системы на web-платформе будет являться некоторая прикладная sc-модель (модель предметной области, закодированная с помощью SC-кода).

В качестве одного их возможных подходов к реализации интеллектуальных систем для web-пространства на основе sc-модели предлагается использовать подход на основе иерархической модели семантического web-сайта (где семантическим web-сайтом будем называть сайт, построенный с использованием средств SC-кода). Основным принципом этой модели является интерпретация прикладной sc-модели на различные традиционные прикладные инструментарии (Рис. 1) для разработки web-сайтов согласно ранее выделенным выше уровням.

Основной задачей при таком подходе является четкая трансляция различных представлений прикладной sc-модели на различных уровнях (уровне внешнего представления, уровне языков разметки, уровне серверных скриптов, уровне хранилищ данных.) для обеспечения тесной интеграции между выделенными уровнями. В sc-системах такой класс задач решает компонент транслятор. Основными языковыми средствами, с которыми будет работать разработчик семантического web-сайта, будут SCnML-тексты.

- операции трансляции SCnML-текстов в некоторое представление в хранилище данных и обратная трансляция этого представления в тексты SCnML;
- операции трансляции SCnML-текстов в эквивалентные объекты SCnML-тегам объекты языков прикладного программирования и обратная трансляция.

2.3. Интерпретация прикладной sc-модели на SCn-модель web-сайта

Основной задачей при интерпретации прикладной sc-модели на sc.n-модель web-сайта является выявление наиболее важных сущностей с точки зрения подачи информации пользователю web-сайта. Такие сущности, как правило, соответствуют некоторым web-страницам в традиционных web-сайтах, согласно базовым принципам представления гипертекстов. В рамках данной работы для описания сущностей предметной области мы используем sc.n-статьи. Кроме определения сущностей, которые будут описываться с помощью sc.n-текстов, на плечи разработчика ложится задача разработки

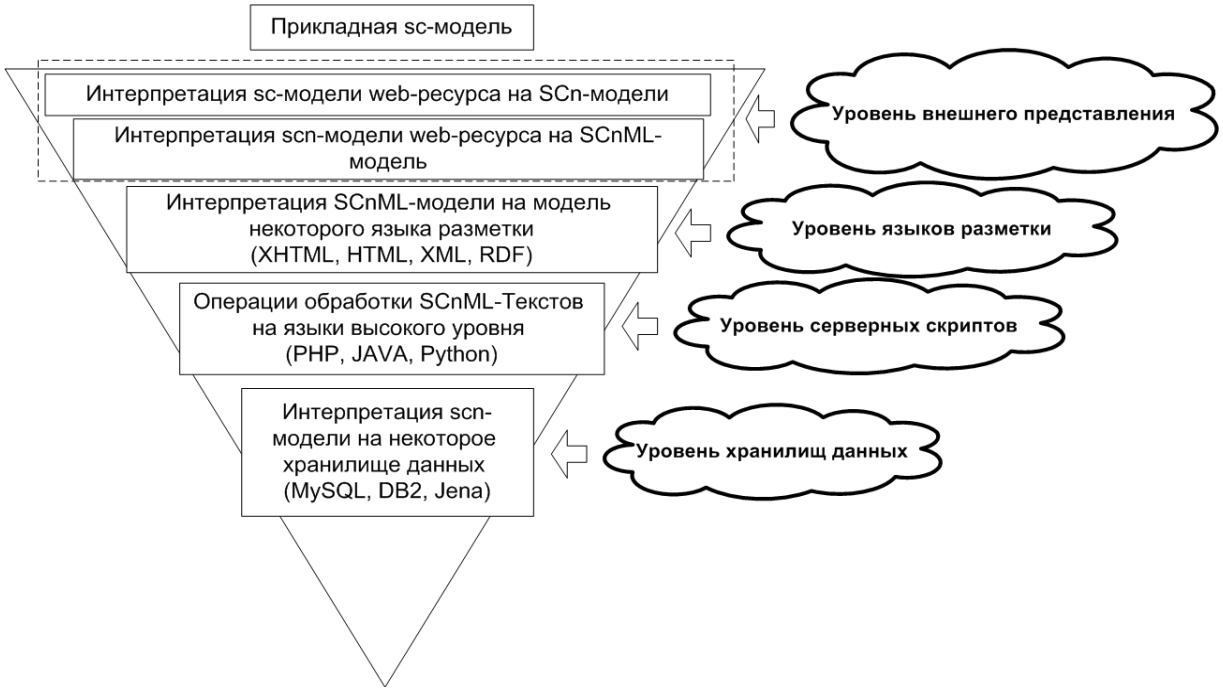

Рисунок 1 – Иерархическая модель семантического web-сайта

Таким образом, можно выделить следующие классы операций трансляции, которые необходимо реализовать при предлагаемом подходе:

- операции трансляции SCnML-текстов в тексты внешнего sc.n-представления;
- операции трансляции SCnML-текстов в тексты языков разметки, используемых в web-ресурсах(HTML, XHTML, XML, RDF и др.);
- операции трансляции SCnML-текстов в эквивалентные объекты SCnML-тегам объекты скриптовых языков уровня разметки (JavaScript) и обратная трансляция;

логической структуры web-страниц, которые будут содержать некоторое множество таких сущностей. В трактовке, предлагаемой в рамках данной работы, web-страница – это методологическая единица, которая определяет порядок и приоритеты пользователя при ознакомлении с материалами, расположенными на странице сайта. Такие web-страницы мы будем называть sc.n-страницами. Таким образом, результатом интерпретации прикладной sc-модели на sc.n-модель web-сайта будет являться множество sc.n-страниц,

соответствующее некоторой методике изложения материала web-сайта.

2.4. Интерпретация SCn-модели web-ресурса на традиционные языки разметки и языки Semantic Web

Для полной интеграции с традиционными инструментариями разработки web-сайтов необходим прозрачный и понятный механизм интеграции, который позволяет мощность традиционных средств разработки сочетать с формальной точностью семантического представления информации. Для этих целей в рамках данной работы предлагается использовать модель языковых средств разметки SCn-кода, SCnML.

Alice
⟵ *foaf:Person*
⟹ *foaf:knows*
- *Eve*
  - ⟹ *foaf:name* [“Eve”]
  - ⟹ *foaf:homepage* [“http:/example.com/Eve”]
  - ⟵ *foaf:Person*
- *Manu*
  - ⟹ *foaf:name* [“Manu”]
  - ⟹ *foaf:homepage* [“http:/example.com/Manu”]
  - ⟵ *foaf:Person*
- *Bob*
  - ⟹ *foaf:name* [“Bob”]
  - ⟹ *foaf:homepage* [“http:/example.com/Bob”]
  - ⟵ *foaf:Person*
  - ⟹ *dc:creator* *Beatufil Sunset*
    - ⟹ *dc:title* [“Beatufil Sunset”]

Рисунок 2 “Социальная сеть Алисы” в виде sc.n-статьи.

Для интерпретации sc.n-текстов на одну из реализаций scnml-разметки разработчику семантического сайта необходимо выполнить следующие действия: разработать способ представления различных типов связок отношений и разработать способ представления различных типов компонент данных связок.

Рассмотрим такую интерпретацию SCn-текстов (Рис. 2) на примере описания “Социальной сети Алисы”, который является одним из известных примеров применения RDF-модели при разметке web-документов с помощью языков HTML или XHTML[RDFa primer, 2011]. Пример выбран не случайно и позволяет на наш взгляд показать, как можно использовать технологии на базе SCn-кода и sc-моделей при разработке web-сайтов.

На основании представленного, на рисунке 2 sc.n-текста мы выделяем основной объект описания ***Alice*** и выделяем следующие типы связок отношений:

- многокомпонетные типы связок: foaf:knows, foaf:homepage, foaf:homepage, dc:creator, dc:title;
- однокомпонетные типы связок: foaf:Person.

Также выделим следующие компоненты связок: Eve, [“Eve”], [“http:/example.com/Eve”], [“Manu”], [“http:/example.com/Manu”], [“Bob”], [“http:/example.com/Bob”], Beatufil Sunset, [“Beatufil Sunset”]. Все выделенные компоненты связок будут являться компонентами связок без вложенного ролевого отношения.

Введем идентификатор сущности ***Alice*** *"http://example.com/alice#me"* этот идентификатор используется в атрибуте RDFa ***about*** для указания того, что вложенные в HTML тег элементы описывают сущность, указанную в значении свойства ***about.*** В начале sc.n-статьи у нас указанно, что объект ***Alica*** является элементом множества ***foaf:Person*** (этот факт sc.n-статье показывает специальный маркер “⟵”). На основании выше сказанного фрагмент статьи, описывающий первые две строки sc.n-статьи будут выглядеть следующим образом:

```
<div class = "scnarticle" about="http://example.com/alice#me">
  <div class = "concept" about="http://example.com/alice#me">
      <div typeof="foaf:Person" >
          <div property="foaf:name">Alice</div>
       </div>

…
</div>
```

Перейдем к описанию части sc.n-статьи, содержащей бинарное отношение ***foaf:knows*** (о том, что отношение бинарное в sc.n-статье показывает специальный маркер “⟹”). Ввиду однотипности связанных отношением ***foaf:knows*** объектов покажем, как представляется нижняя часть sc.n-статьи (рис. 2 выделенный штриховым контуром).

В RDFa информация о связи некоторого web-ресурса или фрагмента web-документа с другим ресурсом или фрагментом web-документа, который описывается в статье, указывается в атрибуте ***rel.*** Атрибут ***property*** позволяет указать, каким отношением текстовый фрагмент в web-документе связан с описываемым объектом. Таким образом, получается, что бинарные ориентированные отношения ***foaf:name***, ***foaf:homepage***, ***dc:title*** мы будем оформлять с помощью атрибута ***property,*** так как данные отношения используются для связи объекта ***Bob*** с некоторыми фактами о нём, представленными в текстовом виде прямо в документе. Отношение ***dc:creator*** мы будем

оформлять с помощью атрибута ***rel,*** так как оно связано фрагментом web-документа. С учетом приведенных выше уточнений, sc.n-статья будут выглядеть следующим образом:

```
…
<div class ="connective" about="http://example.com/alice#me"
rel="foaf:knows">

…
 <div class ="component">
  <div typeof="foaf:Person">
    <div property="foaf:name"> Bob </div>
      <a rel="foaf:homepage" href="http://example.com/bob">
Bob</a>
       <div class="in" about="http://example.com/bob">
         <div class ="connective" rel="dc:creator"
about="http://example.com/bob">
           <div class ="component">
             <div property="dc:title">Beautiful Sunset</div>
           </div>
         </div>
       </div>
       </div>
       </div>
    </div>
```

Атрибуты *class* со значениями *“component”* и *“connective”* в данном фрагменте используются для указания структурных элементов sc.n-статьи. Для уменьшения объёма, где не требуется явного выделения scnml-тегов для реализации верстки sc.n-отображения данной sc.n-статьи (так как, scnml-теги просматриваются, если соотнести sc.n-текст и предложенную в данном примере его html-верстку), структурные элементы sc.n-статьи опущены.

Уточним еще ряд элементов, которые важны при четкой интерпретации. В отличие от RDF модели, в sc-моделях знания об объектах предметной области не привязаны к понятию тройки (субъект предикат объект), поэтому SCn-код не требует в своем представлении элементов подобных ***about***, ***rel***. Кроме этого, необходимо заметить, что, благодаря тому, что sc-модели изначально не привязаны к web-платформе, необходимость в элементах типа ***href***, которые позволяют указать ссылку на некоторых внешний ресурс, отпадает.

Отметим также, что для упрощения примера из текста HTML убраны объявления пространств имен для RDF-словарей “Дублинского ядра” (dc) и FOAF (foaf).

Разобранный пример позволяет продемонстрировать один из возможных подходов к представлению sc.n-текстов с помощью средств HTML или XHTML+RDFa. Отметим, что данный пример приведен лишь с целью продемонстрировать прозрачность перехода от sc.n (или sc-представления) к некоторому более традиционному виду web-ресурса. Для практического использования целесообразнее применять другие, более гибкие подходы, связанные с использованием цепочек преобразований в основе, которых лежит единое формальное представление в виде SC-кода.

2.5. Интерпретация sc.n-модели на хранилища данных

Семантическая структуризация на уровнях внешнего представления и уровне языков разметки обеспечивают, как правило, эффективную навигацию и поиск как в ручном режиме, так и с привлечением сторонних поисковых средств. Однако в современных web-сайтах не вся информация представлена с помощью текстов разметки, большая часть информации генерируется динамически, используя ту информацию, которая находится в некотором хранилище данных. При

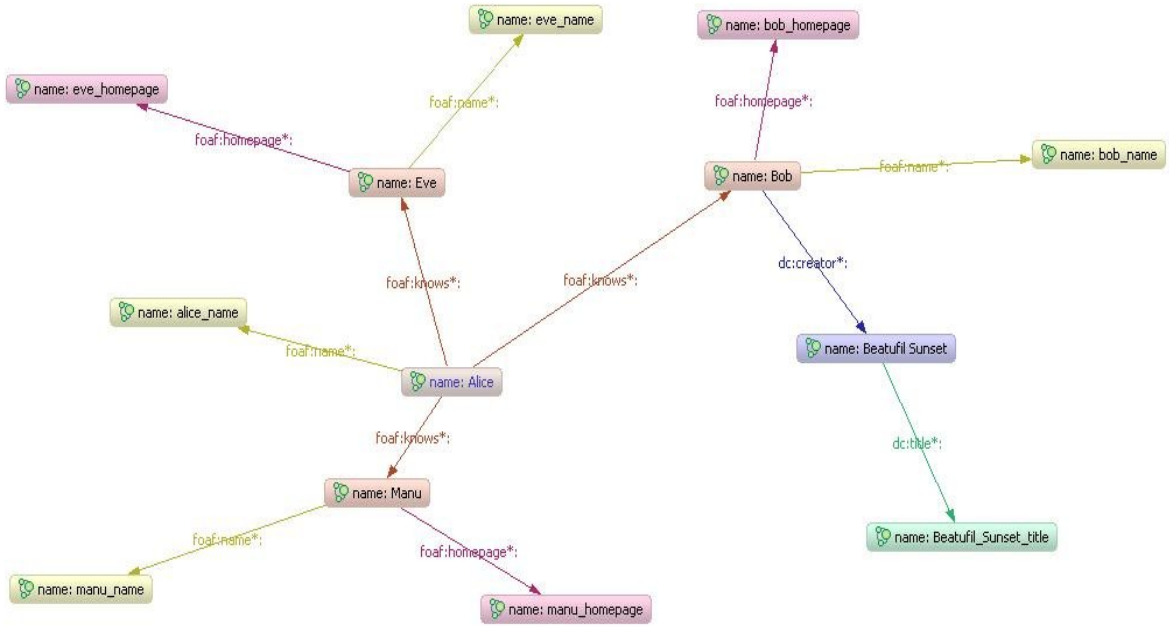

Рисунок 3 Пример sc.n-статьи "Социальная сеть Алисы" представленная в СУБД Neo4j (визуальное представление Neoclipse).

использовании предложенного подхода эффективность такого архитектурного решения будет зависеть в первую очередь от эффективного представления sc-модели или sc.n-статей в рамках некого хранилища данных. В качестве примера интерпретации возьмем интерпретации sc.n-модели web-ресурса на графовую СУБД Neo4j [Neo4j, 2011].

СУБД Neo4j поддерживает модель модифицированной семантической сети.

- связка будет представляться в Neo4j узлом в том случае, если она является связкой неориентированного бинарного отношения или связкой небинарного отношения;
- будем добавлять в конце идентификатора связки терм ":_" в том случае, если она является связкой атрибутивного отношения;
- будем добавлять в конце идентификатора связки терм ":*" в том случае, если она не является связкой атрибутивного отношения.

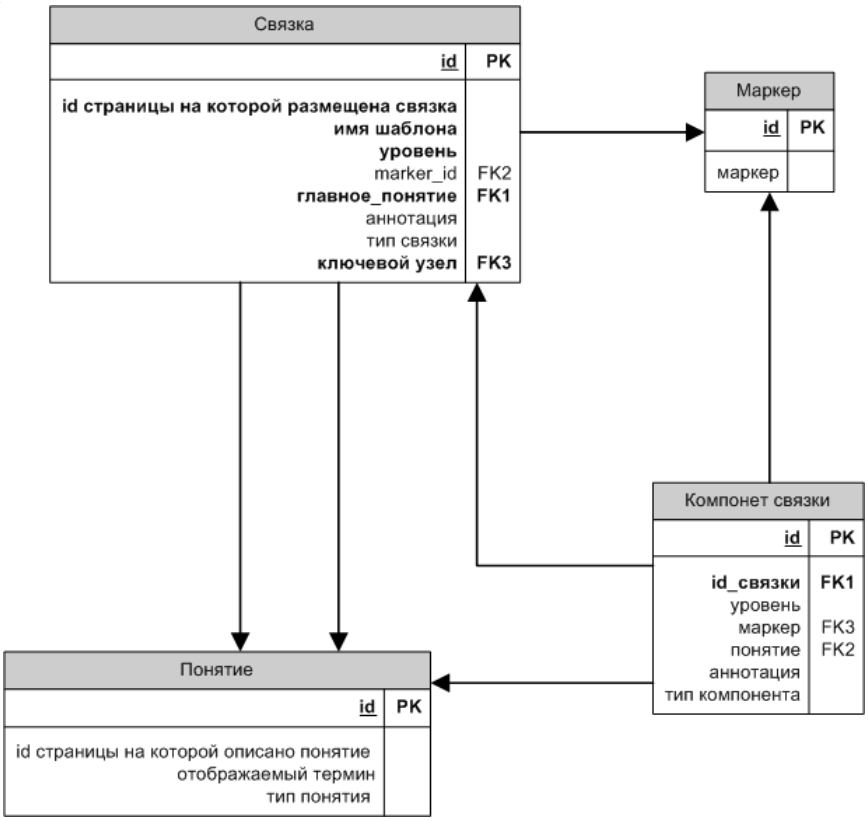

Рисунок 4 Схема экспериментальной базы данных для преставления sc.n-статей на сайте conf.ostis.net

Модификация заключается во введении специальных типов узлов и дуг, которые могут быть проиндексированы. За счет механизма индексов в этой СУБД реализованы эффективные средства для реализации различных алгоритмов поиска на графах.

Существенным отличием SC-кода от других средств представления семантических сетей является возможность проведения sc-дуг из sc-узлов sc-дуг в другие sc-дуги. Поэтому проблему представления таких видов конструкций придется решить для СУБД Neo4j. Введем следующие допущения для СУБД Neo4j:

- связка будет представляться в Neo4j дугой в том случае, если она является связкой ориентированной бинарной пары, и в том случае, если не требуется уточнение атрибутивным отношением роли каждого компонента связки в рамках данной связки;

Введенных допущений достаточно для представления широкого класса sc.n-текстов. Однако для того, чтобы поддержать все возможности SC-кода необходимо будет ввести специальные обозначения для возможных в SC-коде структурных типов sc-элементов.

Введем еще одно допущение, которое позволят нам моделировать sc.n-страницы и sc.n-статьи в рамках Neo4j. Такое допущение вводится в целях реализации эффективной навигации и поиска в рамках семантического web-сайта при предлагаемом нами подходе. Допущение будет заключаться в группировке узлов и дуг, которые встречаются в конкретной sc.n-статье по sc.n-статьям в рамках представления в Neo4j, и группировки sc.n-статей в рамках sc.n-страниц. Такие группировки мы будем осуществлять, используя механизм индексации узлов и связей Neo4j. Благодаря такому механизму у нас с одной стороны сохранится полноценная семантическая

сеть, а с другой стороны будет обеспечена эффективная навигация между структурными элементами семантического web-сайта – sc.n-статьями, sc.n-страницами и sc-элементами, которые в них расположены. Результат интерпретации sc.n-статьи на СУБД Neo4j представлен на рис.3.

Для поддержки sc-файлов, которые используются в SC-коде для представления мультимедиа данных, в Neo4j предусмотрены “property” у узлов и связей. Эти элементы позволяют хранить различные типы данных и допускают бинарное и текстовое представление этих данных. Ввиду того, что Neo4j позволяет эффективно поддержать все элементы SC-кода, можно предположить, что эта СУБД будет являться эффективной платформой для программной реализации sc-хранилищ (хранилищ знаний, закодированных с помощью SC-кода).

Рассмотрим решение, которое позволяет осуществить интерпретацию sc.n-модели web-сайта на реляционные СУБД. Наиболее перспективным, на наш взгляд, подходом отображения любых семантических моделей на реляционные СУБД является способ, в основе которого лежит представление не самих элементов семантической сети, а структурных элементов более высокого уровня абстракции – понятий и отношений. Такое представление хотя и обедняет элементную базу семантической сети, однако позволяет эффективно решать практические задачи.

Для решения задачи представления sc.n-модели web-сайта, на наш взгляд, наиболее простым решением является использование в качестве основных отношений базы данных отношения соответствующие типологии связок и компонентов связок. Для более четкой структуризации необходимо ввести отношения, связывающие набор связок и их компонентов с sc.n-статьей и sc.n-страницей.

В качестве примера интерпретации sc.n-модели web-сайта на реляционную СУБД приведем (рис. 4) схему фрагмента базы данных, которая использовалась для поддержки тегов scnml-запросов в рамках сайта конференций OSTIS. Как видно из рисунка, схема БД существенно облегчена, за счет использования в рамках сайта конференции одной sc.n-страницы одной sc.n-статьи. Из примера понятно, что схема реляционной базы данных при интерпретации sc-моделей может быть разработана с учетом типологии используемых в рамках конкретной реализации семантического web-сайта отношений, их связок и компонентов этих связок.

2.6. Операции обработки scnml-текстов на языках высокого уровня

Уровень серверных скриптов – это часть сайта, которая позволяет обеспечить интерактивное общение с пользователем. При реализации подходов к обработке sc.n и scnml-текстов можно выделить следующие направления работ:

- разработка библиотек для эффективной обработки отдельных фрагментов sc.n-страниц на стороне web-клиента с использованием технологии AJAX;
- разработка серверных библиотек для генерации текстов SCnML-разметки на лету или из хранилища данных;
- разработка серверных библиотек обработки текстов SCnML-разметки, представленных в виде текста;
- разработка библиотек общения с хранилищами данных посредством технологий подобных ORM (Object Relation Mapping) для реляционных СУБД или OTM (Object to Triples Mapping) для RDF-хранилищ;
- разработка прикладных интерфейсов общения между различными семантическими web-сайтами на основе SC-кода, и, как следствие, реализация интерфейсов для общения с поисковыми системами. Такой интерфейс должен обеспечить не только эффективное индексирование текстового представления sc.n-страниц, но и позволить использовать средства семантического поиска на основе специального языка вопросов.

Направления практических работ, которые указаны выше, не являются новыми с точки зрения подходов к разработке традиционных сайтов, однако, реализация указанных инструментов предусматривает реализацию ряда алгоритмов, которые непосредственно связаны с задачами обработки семантических сетей.

ЗАКЛЮЧЕНИЕ

Отметим ряд достоинств предлагаемого подхода к разработке интеллектуальных систем на базе web-платформы:

- подход является полностью независимым от программных средств реализации за счет использования единой формальной основы – sc-модели предметной области, использование таких моделей дает возможность реализации в рамках web-сайтов средств семантического поиска;
- при использовании предложенного подхода web-сайты будут совместимы на уровне SCn и SCnML представлений, что позволяет, используя средства интеграции баз знаний, построенных с использованием SC-кода, объединять семантические web-сайты в семантические web-порталы знаний, которые будут построены по единым унифицированным принципам;
- построенные при использовании предложенного подхода web-сайты можно рассматривать как исходные тексты баз знаний;
- предложенный подход позволяет эффективно использовать технологии Semantic web в качестве платформы интерпретации sc-моделей.

Предложенный подход разработан в рамках проекта OSTIS и опробован на следующих сайтах:

- сайт конференций OSTIS (http://conf.ostis.net);

• информационно справочная система по геометрии (http://ostisgeometry.sourceforge.net/).

Работа выполнена при поддержке гранта БРФФИ Ф10М-085 и гранта БРФФИ-РФФИ Ф10Р-175.

БИБЛИОГРАФИЧЕСКИЙ СПИСОК

[Веб-фрагмент, 2011] Спецификация формата веб-фрагмента – версия 0.9 [Электронный ресурс]. – 2011. - Режим доступа: http://msdn.microsoft.com/ru-ru/library/cc304073%28VS.85%29.aspx . – Дата доступа: 30.11.2011

[Вики о микроформатах, 2011] Вики о микроформатах [Электронный ресурс]. – 2011. - Режим доступа: http://microformats.org/wiki/Main_Page-ru. – Дата доступа: 30.11.2011

[Гаврилова, 2000] Гаврилова, Т.А. Базы знаний интеллектуальных систем/ Т.А. Гаврилова, В.Ф. Хорошевский //СПб – Питер, 2000 г., 384 стр.

[Голенков, 2001] Представление и обработка знаний в графодинамических ассоциативных машинах /В. В. Голенков, [и др]; – Мн. : БГУИР, 2001.

[Грибова, 2011] Грибова, В.В. Облачная платформа для разработки и управления интеллектуальными системами //В. В. Грибова, [и др]; // OSTIS-2011/ Открытые семантические технологии проектирования интеллектуальных систем// Материалы международной научно-технической конференции. - 2011. - С. 5-15.

[Колб, 2009] Колб, Д.Г. Средства просмотра баз знаний интеллектуальных систем / Д. Г. Колб // Вестник БрГТУ. - 2009. - № 5. - С.58-62.

[Колб, 2011] Колб, Д.Г. Направления, методы и средства применения семантических сетей в Internet-технологиях/ Д. Г. Колб // OSTIS-2011/ Открытые семантические технологии проектирования интеллектуальных систем// Материалы международной научно-технической конференции. - 2011. - С.443-463.

[Новожилов 2011] Новожилов, А. Обзор NoSql систем. 2011. - Режим доступа: http://habrahabr.ru/blogs/nosql/77909/ . – Дата доступа: 30.11.2011

[Раскин, 2004] Джеф Раскин Интерфейс: новые направления в развитии компьютерных систем.- Пер. с англ.-СПб: Символ-Плюс, 2004-272 с.

[Хорошевский, 2008] Хорошевский, В.Ф. Пространства знаний в сети Интернет и Semantic Web (Часть 1) / В. Ф. Хорошевский // Искусственный интеллект и принятие решений. - 2008. - № 1. - С.80-97.

[DCMI, 2011] The Dublin Core Metadata Initiative [Электронный ресурс]. – 2011. - Режим доступа: http://dublincore.org/. – Дата доступа: 30.11.2011

[FOAF, 2011] The Friend of a Friend (FOAF) project [Электронный ресурс]. – 2011. - Режим доступа: http://www.foaf-project.org. – Дата доступа: 30.11.2011

[Mercury, 2011] The Mercury Project [Электронный ресурс]. – 2011. - Режим доступа: http://www.mercury.cs.mu.oz.au/index.html. – Дата доступа: 30.11.2011

[Microsoft, 2001] Microsoft Inductive User Interface Guidelines 2011 Microsoft Corporation February 9, 2001 [Электронный ресурс]. – 2011. - Режим доступа: http://msdn.microsoft.com/en-us/library/ms997506.

[Neo4j, 2011] Neo4j the graph database. [Электронный ресурс]. – 2011. - Режим доступа: http://neo4j.org. – Дата доступа: 30.11.2011

[NoSQL, 2011] NoSQL. [Электронный ресурс]. – 2011. - Режим доступа: http://nosql-database.org. – Дата доступа: 30.11.2011

[OSTISa, 2011] Проект 10. Унифицированный способ абстрактного кодирования семантических сетей [Электронный ресурс]. – 2011. - Режим доступа: http://www.ostis.net/wiki/Проект_10. – Дата доступа: 30.11.2011

[RDFa primer, 2011] RDFa Primer Bridging the Human and Data Webs W3C Working Group Note 14 October 2008 [Электронный ресурс]. – 2011. - Режим доступа: http://www.w3.org/TR/2008/NOTE-xhtml-rdfa-primer-20081014/

[Topaz, 2011] Topaz. [Электронный ресурс]. – 2011. - Режим доступа: http://www.topazproject.org. – Дата доступа: 30.11.2011

[W3C, 2007] Multimedia Vocabularies on the Semantic Web W3C Incubator Group Report 24 July 2007 [Электронный ресурс]. – 2011. - Режим доступа: http://www.w3.org/2005/Incubator/mmsem/XGR-vocabularies/#MusicXML

[W3C, 2011] World Wide Web Consortium [Электронный ресурс]. – 2011. - Режим доступа: http://www.w3.org. – Дата доступа: 30.11.2011

THE WEB-BASED IMPLEMENTATION SEMANTIC MODELS OF INTELLIGENT SYSTEMS

Kolb D.G.

Belarusian State University of Informatics and Radioelectronics, Minsk, Republic of Belarus

kolb@bsuir.by

Principles and approaches to development of semantic web-sites with using facilities traditional technoligies are described. The proposed approaches found on SC-model's concept and the way of the pseudonaturaly representation these models.

INTRODUCTION

The most sharpest problems for developers intelligent systems is problems of the development of unified ways and tools advancement systems of artificial intelligent in Internet. The main way for organization traditional informational systems (IS) in the Internet is organization it's in like of web-portal or web-site. For that organization IS can be extract following levels on which can be exercise conversion tradional web-resource into intellectual web-resource. There are levels of external representation, markup languages, server's script and stores

MAIN PART

As the formal basis of the proposed approach will use the semantic network with a base set-theoretic interpretation. The main method of information coding for these networks is the SC (Semantic Code)-code. Intelligent systems built using SC-code is called sc-systems. Entities web-pages will be represent the elements of a knowledge base (KB). This approach allows us to consider a web-site as a specialized intelligent system that solves the problem of dialogue and substantive rights of intellectual systems and providing solutions to key subject of intellectual systems.

Looking at in this perspective, a web-site, we conclude that the degree of intelligence web-site will depend on the extent to which each of the extracted web-site levels of semantic technologies.

CONCLUSION

The main advantage of the introduced approach is the ability to use it in developing intelligent systems for various web-platforms, including Semantic Web platform.