



OSTIS-2013

(Open Semantic Technologies for Intelligent Systems)

УДК 004.822:514

РАЗРАБОТКА WEB-САЙТОВ СО СЛОЖНОСТРУКТУРИРОВАННЫМ КОНТЕНТОМ НА ОСНОВЕ СЕМАНТИЧЕСКИХ СЕТЕЙ

Колб Д.Г., Фурман О.Д.

*Белорусский государственный университет информатики и радиоэлектроники,
г. Минск, Республика Беларусь*

kolb@bsuir.by

olgasharpio@gmail.com

Приведен подход к построению семантически структурированных web-сайтов. Кратко описана модель web-сайтов данного класса, рассмотрена идея организации навигационного поиска в web-сайтах данного класса. Предложен метод отображения класса семантических сетей, используемых для разработки семантически структурированных web-сайтов, на модель RDF.

Ключевые слова: семантическая сеть; интеллектуальная система; Semantic web.

ВВЕДЕНИЕ

Результаты исследования независимых аналитических агентств Gartner Group и Real Story Group за 2011-2012 гг. показывают, что одной из важнейших тенденций развития систем управления web-контентом, в частности в корпоративных web-сайтах, является переход от web-сайтов с простой структурой к крупным интегрированным порталам корпоративных знаний. Указанная тенденция позволяет существенно повысить эффективность использования накопленного корпоративного опыта. С точки зрения разработчика, порталы – это web-сайты, характеризующиеся представлением сложноструктурированной информации и требующие многолетней информационной поддержки. Основной ценностью в них являются накапливаемые за время поддержки объемы корпоративных знаний.

Под сложноструктурированной информацией в данном случае понимаются многоуровневые гипертекстовые информационные структуры, в которых имеют место связи не только между сетями (первичными элементами гипертекста), но и между связями, а также между целыми информационными структурами. Связи в таких иерархических информационных структурах могут быть не только бинарными, но и множественными; не только позитивными, но и негативными и нечеткими; не только стационарными во времени, но и не стационарными; не только константными, но и переменными. Примерами web-сайтов, оперирующих такой информацией, являются

научные, образовательные, энциклопедические порталы, порталы проектных организаций и др. Очевидно, что для разработки web-сайтов указанного класса необходимо использовать модели, позволяющие формально представлять всё многообразие связей, присутствующих на web-сайте. Нам представляется, что наиболее подходящей моделью для представления сложноструктурированной информации является семантическая сеть.

Проблема разработки web-сайтов, в основе которых лежат семантические сети, всё чаще входит в сферу интересов многих ученых всего мира. С одной стороны решение этой проблемы позволит перейти на новый качественный уровень Всемирной Паутины – Semantic Web, с другой стороны позволит преодолеть еще одну ступеньку на пути к созданию интеллектуальных систем и заложить платформу для создания интеллектуальных систем для Internet.

Жизненный цикл процесса разработки web-сайтов включает следующие основные задачи:

- наполнение контента web-сайта (в том числе и мультимедийного);
- оптимизация контента для поисковых машин;
- разработка средств навигации и поиска по web-сайту;
- обеспечение переносимости контента web-сайта.

Современные CMS с различным уровнем полноты решают указанные задачи, однако остается

и ряд не решенных задач, решение которых невозможно получить без использования методов и средств искусственного интеллекта. В число таких задач входят:

- обеспечение высокого уровня релевантности поисковых запросов ответам;
- семантическая структуризация и оптимизация страниц web-сайтов с учетом формальной спецификации, которую может “понять” и проиндексировать поисковая машина;
- обеспечение кроссплатформенности контента web-сайта не только между различными версиями программных средств, а и между различными программными средствами реализации.

В рамках данной работы авторы попытаются рассмотреть возможные пути решения, указанных проблем, на основе средств, которые предлагаются в рамках направления Semantic Web.

1. Модель семантически структурированного web-сайта

1.1. Зачем нужны модели web-сайтов?

Рассматривая решения, которые в настоящее время предлагаются для совершенствования web-ресурсов указанного класса, необходимо отметить два значимых направления работ:

1. разработка методов и средств спецификации страниц web-сайта, которые позволяют осуществлять лёгкую индексацию web-страниц и, соответственно, обеспечить более релевантный поиск;
2. разработка методов и средств спецификации информации, которые позволяют представлять сложноструктурированные виды информации, в том числе, такие как теоремы, аксиомы, логические высказывания и др.

К первому направлению можно отнести ряд работ, позволяющих на основе модификации языка разметки осуществить семантическую разметку web-страницы [Микроформаты, 2012], [Микроданные, 2012], [RDFa, 2012] или её семантическую аннотацию [SMW, 2012], [SMW+, 2012], [KnowWE, 2012]. Средствами этого направления, как правило, присуща малая выразительная мощность [Хорошевский, 2008]. Однако использование этих средств существенно облегчает работу по семантическому анализу содержимого web-страницы, которая в настоящее время чаще всего осуществляется с помощью средств Text mining.

Ко второму направлению относятся работы, в основе которых лежит представление web-ресурсов в виде семантических сетей на основе модели RDF, используя базовые RDF-словари RDFS, OWL или OWL 2. Выразительная мощность указанных средств прежде всего, определяется практической направленностью задач, которые пытаются решать различные научные, участвующие в разработке стандартов направления Semantic Web. С точки

зрения практической используемости для разработки web-сайтов данные средств тяжелы в освоении и далеко не всегда могут удовлетворить требования разработчика интеллектуальных систем для Internet [Хорошевский, 2008].

Возникает дилемма – использовать маловыразительные средства, которые адаптированы для задач промышленной разработки web-приложений или использовать средства, обладающие большей выразительной мощностью, однако не адаптированные для промышленной разработки.

В рамках данной работы авторы попытаются предложить решение, которое позволит интегрировать два приведенных выше направления разработки web-приложений, на основе использования моделей представления знаний более высокого абстрактного уровня. Такой подход с точки зрения авторов позволит с одной получить промежуточное решение для прикладной промышленной разработки web-сайтов на базе семантических сетей. И даст толчок к эволюционированию существующих в настоящее время средств разработки web-приложений, основанных на семантических сетях.

1.2. Компоненты семантической модели web-сайта

Семантически структурированными web-сайтами назовём специальный класс интеллектуальных систем, построенных на базе семантических сетей и гипертекстовой модели представления информации. Модель семантически структурированного web-сайта задается следующим образом:

$$M = \{O, S_A, S_P, O_{hyp}, M_L, M_S, I\}, \quad (1)$$

где

O – онтология предметной области, по которой разрабатывается web-сайт, в рамках которой специфицированы сущности предметной области и отношения между ними;

S_A – множество классов статей семантически структурированного web-сайта (атомарных семантических единиц), которое строится на основании онтологии O предметной области;

S_P – множество классов страниц семантически структурированного web-сайта, каждый из которых состоит из некоторого множества классов статей и каждый элемент, которого связан с другим элементом отношением порядка следования. Такие связи позволяют определить последовательность изложения подачи «материала» предметной области содержимого web-страницы, который будет размещен на web-странице;

O_{hyp} – онтология мультимедийных документов, которая строится с учетом концепции языка гипермедийных сетей, который был разработан для семантической спецификации мультимедийных данных. В его основе лежит универсальная система метаданных «Дублинское ядро» и однородные

семантические сети с базовой теоретико-множественной интерпретацией [Колб, 2009].;

M_L – языковые средства представления информации в виде однородных семантических сетей;

M_S – модель семантического навигационного поиска, которая состоит из семейства операций обработки однородных семантических сетей, обеспечивающих навигацию по семантически структурированному web-сайту;

I – способ отображения модели семантически структурированного web-сайта на модель RDF.

Указанную модель семантически структурированного web-сайта будем называть SC-моделью web-сайта, так как в её основе лежит

предметной области, по которой разрабатывается web-сайт. Класс статей (семантических окрестностей) A_i задается следующим образом:

$$A_i = \{E, S_R, S_{Aij}\}, \quad (2)$$

где

E – сущность предметной области, по которой разрабатывается семантически структурированного web-сайта;

S_R – множество связей отношений, каждое из которых специфицировано в онтологии O таким образом, что сущность E является первым компонентом связки данного отношения;

S_{Aij} – множество статей частного вида, связанных с

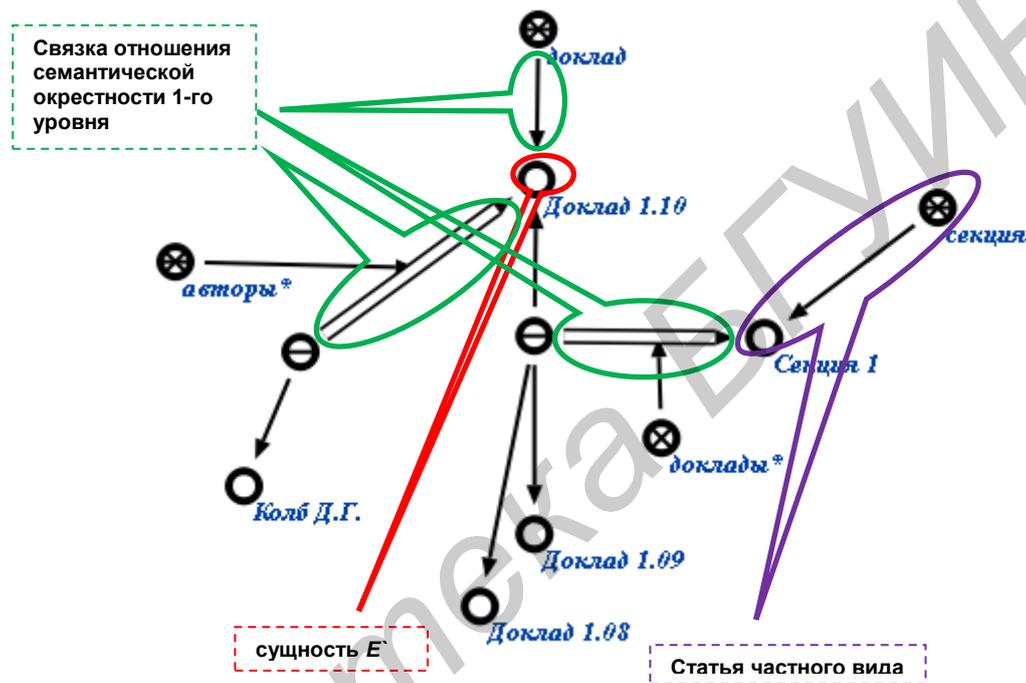


Рисунок 1 – Семантическая окрестность понятия Доклад 1.10

универсальный базовый способ представления семантических сетей в виде однородных семантических сетей с базовой теоретико-множественной семантической интерпретацией. Такой универсальный способ представления семантических сетей описан в работе [Голенков, 2011] и назван SC-кодом (Semantic Code).

Информационные конструкции, представленные в SC-коде, называются SC-графами. Кроме SC-кода, вводится способ символической визуализации SC-графов названный SCn-кодом (Semantic Code natural) [Голенков, 2011]. Текст SCn-кода представляет собой совокупность SCn-статей, каждая из которых является описанием семантической окрестности (см. рисунок 1) некоторого соответствующего ей элемента SC-графа.

Каждый элемент множества S_A задает класс атомарных семантических единиц (статей) семантически структурированного web-сайта. В рамках данной работы статьёй семантически структурированного web-сайта назовем семантическую окрестность некоторой сущности E

сущностью E посредством множества связей отношений S_R .

Каждая семантическая окрестность A_i задается на множестве отношений, связки которых являются элементами S_R , и может характеризовать сущность E с различной степенью глубины. Семантическую окрестность, характеризующую сущность E с глубиной n , назовем семантической окрестностью уровня n . Полной семантической окрестностью сущности E будем называть такую семантическую окрестность, которая включает связи всех отношений, в которые включен знак данной сущности.

Рассмотрим пример формального представления статьи семантически структурированного web-сайта, выделив её элементы согласно (2). Пусть E' – некоторое понятие, которое описывается в рамках статьи A_i' . На рисунке 1 с помощью средств SCg-кода [Голенков, 2011] представлена семантическая окрестность второго уровня для сущности E' – Доклад 1.10. Множество связей отношений, которые являются элементами S_R' в рамках статьи A_i' на рисунке 1 подписаны указателем “связка

отношения семантической окрестности 1-го уровня”. Множество статей частного вида S_{A_i} в рамках статьи A_i на рисунке 1 подписаны указателем “статья частного вида”, здесь у данного множества один элемент соответствующий семантической окрестности сущности **Секция 1 A_0** . Указанная сущность связана с сущностью **Доклад 1.10** через связку отношения **доклады***. A_0 имеет непустое множество S_{0K} . Остальные сущности присутствующие на рисунке (**доклад**, **Доклад 1.09**, **Доклад 1.08**, **Колб Д.Г.**) в рамках данной статьи связаны только с сущностью E и, соответственно, не образуют статей частного вида. Используя конструкции, разобранные в данном примере, мы можем характеризовать любую сущность предметной области с «уровнем подробности», который определяет разработчик web-сайта исходя из своих потребностей. В рамках нашей работы этот «уровень подробности» мы называем семантической окрестностью понятия уровня n .

1.3. Модель семантического навигационного поиска

В семантически структурированных web-сайтах каждый переход между страницами является переходом от одной семантической окрестности некоторой сущности к другой семантической окрестности. Сам переход приводит к выполнению поисковой операции, которая инициируется на основании семантической спецификации операции, формируемой разработчиком web-сайта. Такую операцию можно расценивать как решение некоторой задачи семантического информационного поиска. Решение этой задачи может приводить к инициированию как обычных поисковых операций, так и операций логического вывода. Приведенная идея организации навигационного поиска характерна для интеллектуальных систем, и хотя она высказывалась многими авторами, не поддерживается в поисковых моделях, используемых в Semantic web и являющихся наиболее развитыми моделями семантического поиска.

Наиболее распространённые модели поиска Semantic Web используют для семантических запросов конечный набор операций, которые работают только на основе графовых шаблонов поиска. Такие шаблоны неплохо работают на небольших семантических сетях, на которых легко задать конфигурацию искомого графа. Однако в web-сайтах уровня корпорации, объёмы информации таковы, что возрастает необходимость в сложных запросах, в которых количество элементов шаблона может определяться десятками тысяч элементов. Поэтому появляется необходимость принципиально новых подходов к организации поиска на семантических сетях.

В настоящее время одним из популярных решений указанных выше трудностей является совершенствование языка SPARQL, предложенного сообществом Semantic web, в направлении

расширения его синтаксиса таким образом, чтобы описывать графовый шаблон поиска на основании регулярного выражения или некоторой графовой грамматики [Anyanwu, 2007], [Alkhateeb, 2009], [Pérez, 2010]. Однако такой подход предполагает, что разработчик должен знать топологию семантической сети. Данная информация не всегда бывает доступна разработчику интеллектуальной системы. Поэтому нам кажется, что для построения навигационной поисковой модели для семантически структурированных web-сайтов наиболее эффективен подход, который используется в вопросно-ответных системах [Сулейманов, 2001]. Такой подход дает возможность строить машину поиска как семейство поисковых операций, которые работают с семантической спецификацией вопроса (запроса) пользователя. Типология операций не ограничивается операциями поиска по графовому шаблону, а может быть расширена до любых классов операций, например, таких как операции логического вывода.

Модель навигационного поиска для семантически структурированных web-сайтов, построенных на основе модели (1), зададим следующим образом:

$$M_s = \{L_q, L_a, W\}, \quad (3)$$

где

L_q – язык вопросов, с помощью которого в виде формальных текстов осуществляется описание (спецификация) запрашиваемых (искомых) фрагментов семантической сети;

L_a – язык представления ответов, с помощью которого осуществляется выделение формальных текстов, являющихся ответами, и описание их связи с формальными текстами, которые представляют вопросы, соответствующие указанным ответам;

W – семейство информационно-поисковых операций, каждая из которых реагирует на соответствующий ей семантический тип вопроса и выполняет соответствующую ей поисковую процедуру на семантической сети.

Модель информационного поиска (3) относится к классу моделей вопросно-ответного поиска и позволяет обеспечить семантически структурированный web-сайт базовыми средствами навигационного поиска.

В основе языка вопросов L_q поисковой модели (3) предлагается использовать формальную теорию вопросов [Белнап, 1981]. На этой теории для предлагаемой в работе поисковой модели разработана типология вопросов, которая определяет семантическую мощность предлагаемых средств поиска и задает базовые навигационные операции поисковой модели.

В типологии вопросов выделены следующие типы вопросов:

- *запрос всех элементов множества*: чаще всего заданное множество – это конечное множество из элементов некоторой структуры;

- *запрос внешней информационной конструкции*, представленной некоторым файлом в том или ином формате;

- *что-это-вопрос*, запрашивающий основные сведения об указываемой сущности – семантической окрестности, "центром" которой является знак указываемой сущности.

В рамках данной модели поиска для каждого типа вопроса можно поставить в соответствие некоторый тип ответа. Такие типы ответов определяются структурой представления ответа и видами знаний, которые в них представлены. В общем случае между типами вопросов и ответов нет взаимно-однозначного соответствия. Это означает, что на один и тот же вопрос могут быть получены различные типы ответов.

В современных вопросно-ответных системах принято, что ответ на заданный вопрос формулируется в "естественно-языковой манере вопроса" [Сулейманов, 2001]. В рамках предлагаемой модели эта идея может быть выражена формой представления ответа на некотором внешнем языке, который понятен пользователю семантически структурированного web-сайта и используется для представления информации на его страницах.

Поисковая машина начинает работать в процессе сёрфинга пользователя семантически структурированного web-сайта. Каждый переход по ссылке к другой странице, инициирует операцию, которая работает в соответствии со следующим алгоритмом:

Шаг 1. Анализируется вопрос, поступивший на вход поисковой системе. Вопрос представлен на языке вопросов, который позволяет осуществить его семантическую спецификацию. Анализ семантической спецификации позволяет выявить следующие характеристики вопроса: тип задаваемого вопроса, наличие ответа на данный вопрос (если данный вопрос уже формулировался прежде), входные данные для поисковой операции, которая соответствует вопросу данного класса;

Шаг 2. На основе семантической спецификации, в которую входят входные данные для операции, информация о типе операции, происходит выбор операции из семейства операций поисковой машины. После поиска операции генерируется условие инициирования операции

Шаг 3. Выполнение поисковой операции и формировании ответа на поставленный вопрос.

Шаг 4. Чистка семантической сети от сгенерированных поисковой операцией вспомогательных информационных конструкций.

Шаг 5. Уточнение типа ответа для поставленного перед поисковой машиной вопроса, для этого анализируется семантическая спецификация вопроса, если информации о предпочтительном классе ответа нет, то используется прямой ответ.

Шаг 6. Формирование визуального представления вопроса, то есть генерация web-

статьи содержащей ответ на вопрос.

Семейство информационно-поисковых операций W задает набор процедур поиска в соответствии с набором типов вопросов и для базовой модели семантического информационного поиска выглядит следующим образом:

- *операция поиска всех элементов заданного конечного множества* (чаще всего – это множество из элементов некоторой структуры, статей SC-сайта одного класса);

- *операция поиска внешней информационной конструкции*, представленной некоторым файлом в том или ином формате;

- *операция поиска семантической окрестности сущности*, запрашивающая основные сведения об указываемой сущности – семантической окрестности, "центром" которой является знак указываемой сущности.

Каждая из указанных операций описывается на основании графового шаблона поиска, который специфицируется с помощью средств SC-кода.

2. Связь модели web-сайта с платформой Semantic web

Для разметки страниц web-сайта в соответствии с приведенной моделью языков разметки, которые используются в настоящее время не достаточно. В частности представление информации на страницах web-сайта с использованием модели RDF зачастую является сложным и избыточным [Хорошевский, 2011]. А для класса web-сайтов, которые используются в рамках данной работы, не обеспечивает необходимой выразительной мощности. Исходя из приведенных соображений, предлагается использовать язык разметки семантически структурированных гипертекстов.

2.1. Язык разметки семантически-структурированных гипертекстов для представления web-страниц

В качестве языка разметки гипертекстов, в основе которых лежат семантические сети, предлагается использовать SCnML (SCn Markup Language) – язык разметки семантически структурированных гипертекстов [Голенков, 2011], который задается следующим образом:

$$M_L = \{T_e, T_c, T_{comp}, T_q, R\},$$

где

T_e , – множество *корневых тегов описываемых сущностей* – множество текстовых обозначений сущностей предметной области. Описание каждой такой корневой сущности представляет собой SCnML-текст, соответствующий некоторой SCn-статье.

T_c – множество *тегов связей* – множество текстовых представлений обозначений связей отношений, используемых в рамках данной SCn-статьи. Такие связки отношений соединяют сущности, описываемые в рамках SCn-статьи с

другими сущностями, обозначения которых представлены в SCn-статье. Множество T_c задается следующим образом:

$$T_c = \{T_{c1}, T_{c2}\}$$

T_{c1} – множество **тегов однокомпонентных связей** – множество текстовых представлений связей отношений, каждая из которых в рамках SCn-статьи всегда связывает сущность с одной и только одной сущностью, обозначение которой присутствует в рамках данной SCn-статьи. Множество T_{c1} задается следующим образом:

$$T_{c1} = \{T_{c1s}, T_{c1r}\},$$

где

T_{c1s} – множество тегов однокомпонентных связей без указанной роли компонента связки отношения, которую они играют в рамках данной связки отношения.

T_{c1r} – множество тегов однокомпонентных связей с указанной ролью компонента связки отношения, которую они играют в рамках данной связки отношения.

T_{c2} – множество **тегов многокомпонентных связей** – множество текстовых представлений связей отношений, каждая из которых в рамках SCn-статьи всегда связывает сущность с одной или более сущностями, обозначения которых присутствуют в рамках данной SCn-статьи.

T_{comp} – множество **тегов компонентов связей** – множество текстовых представлений компонентов отношений, каждый из которых в рамках SCn-статьи является обозначением сущности, связанной с корневой сущностью отношением, тип которого задается соответствующим тегом многокомпонентной связки. Множество T_{comp} задается следующим образом:

$$T_{comp} = \{T_{comp1}, T_{comp2}\}$$

T_{comp1} – множество **тегов компонентов связей без указания роли компонента связки** – множество текстовых представлений знаков сущностей, связанных с сущностями, встречаемыми в рамках SCn-статьи.

T_{comp2} – множество **тегов компонентов связей с указанием роли каждого компонента** – множество текстовых представлений компонентов связей отношений, которые являются обозначением сущностей, связанных с некоторыми сущностями, встречаемыми в рамках SCn-статьи, каждый компонент, связки которого уточняется дополнительно ролью.

T_q – множество **тегов SCnML-запроса** – множество текстовых представлений спецификаций семантических запросов к web-ресурсу, разработанному с помощью семантически структурированных гипертекстов (см. листинг ниже и рисунок 2). Использование таких тегов в SCnML-текстах позволяет исключить дублирование информации в рамках различных SCnML-текстах за счет того, что результаты семантических запросов, задаваемых с помощью данного тега, включаются в

SCn-статью и, тем самым, позволяют расширить спецификацию сущности, описываемой в рамках SCn-статьи.

R – множество правил размещения тегов при записи SCnML-текстов. На множестве SCnML-тегов заданы следующие типы отношений: **быть корневым тегом***, **быть родительским тегом***. Введены следующие правила формирования SCnML-текстов:

- каждый SCnML-тег связан с другим SCnML-тегом понятием уровня. В рамках данной работы уровень показывает «через сколько отношений» связана корневая сущность с некоторой другой сущностью в рамках данной статьи. Уровень позволяет задать отношение между родительским и дочерним тегом и определяет, для какой сущности определены связки отношений, расположенные на один уровень ниже, и к какой

OSTIS-2013

=> Конференция OSTIS-2013

= международная конференция

= OSTIS-2012

§ Время проведения:

• с_ :2013.02.21

• по_ :2013.02.23

§ Программный комитет:

- Гаврилова Т. А.
- Голенков В. В.
- Гулякина Н. А.
- Грибова В. В.
- Ефименко И. В.
- Заболеева-Зотова А. В.
- Загоруйко Ю. А.
- Иванюк А. А.

полный состав программного комитета...

§ Пункты питания участников:

- Ресторан "ДжоМалунама"
- Ресторан-бистро "Лидо"
- Кафе "Жбан"
- Столовая "Амкодор"

Результат работы запроса поиска элементов множества программный комитет -2013

Рисунок 2 – SCn-статья, сформированная системой.

связке относятся сущности, являющиеся компонентами связей;

- на одной web-странице могут располагаться SCnML-тексты, соответствующим нескольким SCn-статьям;

- SCnML-текст, соответствующий одной SCn-статье может входить в состав SCnML-текста, соответствующего другой SCn-статье;

- теги в рамках SCnML-текста записываются по следующим правилам:

- первым следует корневой тег, тег корневой сущности;

- за корневым тегом может следовать только тег связки;

- тегу компонента связки всегда предшествует тег связки.

Каждой SCn-статье в рамках семантически структурированного web-сайта соответствует некоторый SC-граф, семантически эквивалентный ей. Ввиду того, что SCnML – это средство разметки SCn-статей, любому SCnML-тексту также соответствует некоторый SC-граф.

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.

Листинг, приведенный ниже, представляет фрагмент SCnML-текста, реализованный с помощью средств wiki-разметки и соответствующий фрагменту SCn-статьи, использованной при спецификации сущности “Конференция OSTIS-2013”.

```

{{SCnFieldConcept|OSTIS-2013}}
{{SCnFieldSpecConSyn|1|Конференция OSTIS-2013}}
{{SCnFieldSpecConMemberSet|1|международная конференция}}
{{SCnFieldSpecConMemberEl|2|OSTIS-2012}}
{{SCnFieldSpecConDate|1}}
{{SCnFieldCompEnumWithAttr|2|1|с|2013.02.21|SCnFieldSpecConDateBegin}}
{{SCnFieldCompEnumWithAttr|2|1|по|2013.02.23|SCnFieldSpecConDateEnd}}
{{SCnFieldSpecConProgramCommittee|1}}
{{SCnFieldCompQueryEnum|2|программный комитет OSTIS-2013|полный состав программного комитета ...|SCnFieldSpecConProgramCommittee}}
{{SCnFieldSpecConFoodItem|1}}
{{SCnFieldCompEnum|2|1|Ресторан "ДжоМалунгма"|SCnFieldSpecConFoodItem}}
{{SCnFieldCompEnum|2|1|Ресторан-бистро "Лидо"|SCnFieldSpecConFoodItem}}
{{SCnFieldCompEnum|2|1|Кафе}}

```

```

"Жбан"|SCnFieldSpecConFoodItem}}
{{SCnFieldCompEnum|2|1|Столовая
"Амкодор"|SCnFieldSpecConFoodItem}}

```

В листинге с помощью тега *SCnFieldConcept* обозначается корневая сущность. Теги в листинге, начинающиеся с префикса *SCnFieldSpecCon*, являются тегами связей соответствующих отношений: *SCnFieldSpecConSyn* – тег однокомпонентной связи отношения синонимии; *SCnFieldSpecConMemberSet* – тег однокомпонентной связи отношения “быть элементом множества”; *SCnFieldSpecConDate* – тег многокомпонентной связи отношения “время проведения”; *SCnFieldSpecConProgramCommittee* – тег многокомпонентной связи отношения “программный комитет”; *SCnFieldSpecConFoodItem* – тег многокомпонентной связи отношения “пункты питания”. Теги, начинающиеся с префикса *SCnFieldComp*, являются тегами компонентов связей соответствующих отношений: *SCnFieldCompEnumWithAttr* – тег компонента многокомпонентной связи, с возможностью указания роли каждого компонента; *SCnFieldCompQueryEnum* – тег SCnML-запроса, результат которого будет выдаваться в виде набора тегов компонентов связей без указания роли каждого компонента; *SCnFieldCompEnum* – тег компонента многокомпонентной связи. У каждого тега связи и тега компонента связи указывается атрибут, являющийся номером уровня. В рамках данного примера номер уровня указывается вторым атрибутом, нумерация начинается с единицы.

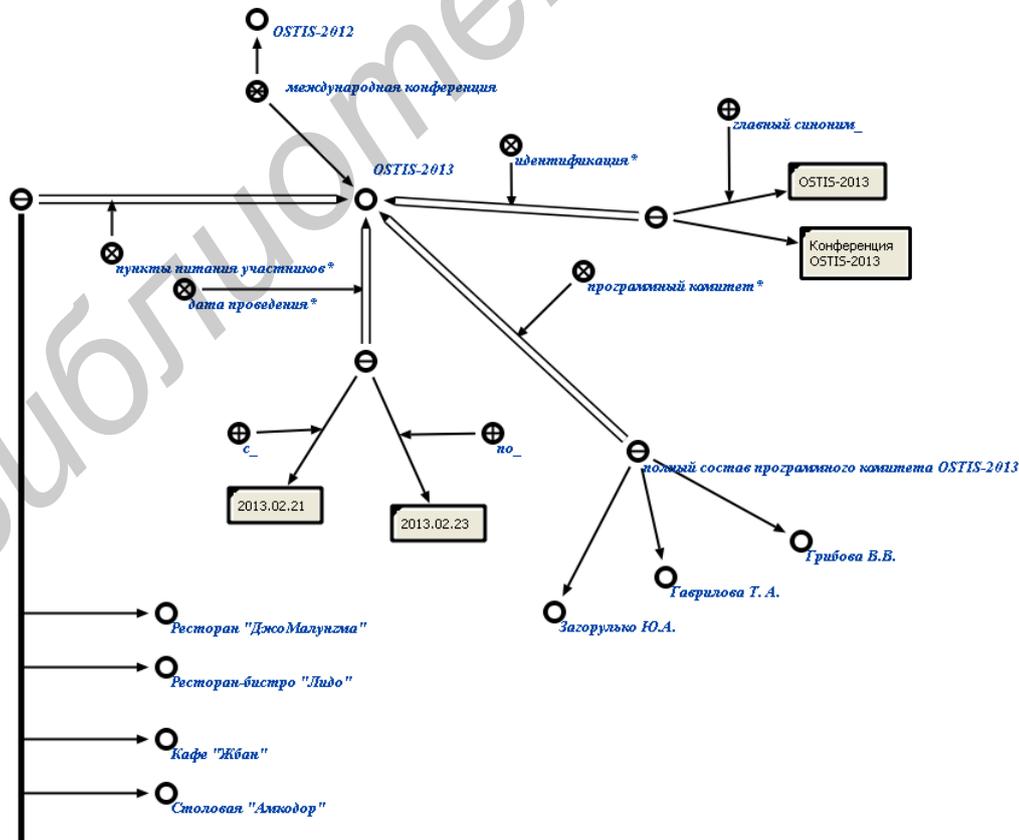


Рисунок 3 – SC-граф, который задает рассматриваемый листинг

После обработки данного листинга инструментальными средствами, на базе платформы Mediawiki, будет сгенерирована SCn-страница, приведенная на рисунке 2.

Каждой однокомпонентной связке соответствует обозначение на SCn-странице специальным маркером. Все многокомпонентные связи отображаются одним маркером и различаются лишь идентификатором отношения, экземпляр которого они обозначают. Роли компонентов связей указываются через двоеточие перед соответствующим компонентом связи, при этом каждое обозначение роли дополнено знаком “_”. Как будет выглядеть SC-граф рассматриваемого примера показано на рис. 3.

2.2. Преобразование семантически-структурированных гипертекстов в RDF-графы

В основе метода преобразования лежит представление элементов алфавита SC-кода с помощью одного из базовых словарей RDF в рамках некоторого инструментального средства, поддерживающего RDF. Метод отображения SC-моделей на RDF-модель I_{SC} задается следующим образом:

$$I_{SC} = \{S_{SC}, ARDF, R\},$$

где

S_{SC} – множество SC-текстов прикладной SC-модели некоторого web-сайта. Такие SC-тексты могут включать, в том числе и тексты программ обработки SC-текстов, представленных средствами SC-кода.

$ARDF$ – множество SC-элементов, представленных в терминах средств метаописаний Semantic Web – RDFS, OWL или OWL2.

R – множество правил перехода от SC-текстов к RDF-текстам. Такие правила оформляются в соответствии с некоторым синтаксическим представлением как SC-текстов, так и RDF-текстов.

Для каждого элемента алфавита SC-кода составляется правило преобразования. Приведем примеры таких правил (таблица 1) для некоторых элементов алфавита SC-кода, которые характеризуют особенности представления данного способа кодирования семантических сетей в RDF.

Из таблицы 1 видно, что для обозначения как sc-узлов, так и sc-дуг в RDF-графе используются узлы. Такой подход обусловлен непригодностью RDF-модели к представлению многоместных отношений и связей между узлами [Трофимов, 2011]. Преобразование на основе указанных в таблице 1 правил дает возможность хранить тексты SC-кода в современных RDF-хранилищах.

В качестве базовых предикатов для представления в RDF семантических сетей в SC-коде введены две связи отношения инцидентности – *компонент_sc-коннектора*, *второй_компонент_sc-коннектора*; в качестве

Таблица 1 – Правила преобразования элементов алфавита SC-кода в RDF-представление с помощью RDFS (нотация N3)

<p>Элемент узлового типа (sc-узел)</p> <pre>@prefix: <http://www.ostis.net/> . @prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> . @prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> . :sc-элемент a rdfs:Class . :sc-узел a rdfs:Class; rdfs:subClassOf :sc-элемент . :xi rdfs:subClassOf :sc-узел.</pre>
<p>Элемент дугового неориентированного типа (sc-ребро)</p> <pre>@prefix: <http://www.ostis.net/> . @prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> . @prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> . :sc-элемент a rdfs:Class . :sc-коннектор a rdfs:Class; rdfs:subClassOf :sc-элемент . :sc-ребро a rdfs:Class; rdfs:subClassOf :sc-коннектор . :компонент_sc-коннектора a rdf:Property; rdfs:domain :sc-элемент; rdfs:range :sc-коннектор . :e rdfs:subClassOf :sc-ребро. :xi :компонент_sc-коннектора :e. :xj :компонент_sc-коннектора :e.</pre>
<p>Элемент дугового ориентированного типа (sc-дуга общего вида)</p> <pre>@prefix: <http://www.ostis.net/> . @prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> . @prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> . :sc-элемент a rdfs:Class . :sc-коннектор a rdfs:Class; rdfs:subClassOf :sc-элемент . :sc-дуга_общего_вида a rdfs:Class; rdfs:subClassOf :sc-коннектор . :компонент_sc-коннектора a rdf:Property; rdfs:domain :sc-элемент; rdfs:range :sc-коннектор . :второй_компонент_sc-коннектора a rdf:Property; rdfs:domain :sc-элемент; rdfs:range :sc-коннектор . :e rdfs:subClassOf :sc-дуга_общего_вида. :xi :компонент_sc-коннектора :e. :xj :второй_компонент_sc-коннектора :e.</pre>

базовых классов элементов семантической сети введены классы RDF-сущностей, которые соответствуют элементам алфавита SC-кода – *sc-элемент*, *sc-коннектор*, *sc-узел*, *sc-ссылка*, *sc-коннектор*, *sc-ребро*, *sc-дуга общего вида*. Для обеспечения унифицированного представления текстов SC-кода на основе структурно-семантической типологии элементов SC-кода и на основе RDFS-словаря разработана онтология элементов SC-кода.

Учитывая сказанное, алгоритм отображения прикладной SC-модели на RDF-модель в соответствии с предлагаемым методом выглядит следующим образом:

Шаг 1. Описание прикладной модели семантически структурированного web-сайта на каком-либо из SC-языков.

Шаг 2. Выбор одного из базовых RDF-словарей для отображения на платформу Semantic Web элементов SC-кода (уточнение какой из RDF-словарей мы будем использовать, RDFS, OWL или OWL 2). Выбор такого словаря определяется практической целесообразностью.

Шаг 3. Определение предметных словарей, которые будут использованы для представления прикладной модели.

Шаг 4. Реализация правил перехода от выбранного синтаксического представления SC-текстов к выбранному синтаксическому представлению RDF-текстов.

Шаг 5. Тестирование прикладной RDF-модели семантически структурированного web-сайта.

3. Апробация модели семантически структурированного web-сайта

На основании анализа современных инструментариев, используемых при построении семантических систем, был сформулирован ряд ключевых принципов разработки семантически структурированных web-сайтов:

- процесс разработки интеллектуальных систем – процесс коллективный и требующий постоянного согласования, поэтому должна быть использована гибкая и расширяемая платформа, ориентированная на коллективную работу;

- процесс разработки интеллектуальных систем должен протекать открыто и к такому процессу должен привлекаться как можно более широкий круг участников. Поэтому используемые средства должны быть открытыми и легко доступными;

- платформа разработки семантических и интеллектуальных систем должна обеспечивать простой интерфейс для подключения механизмов удаленного управления контентом для того, чтобы была возможность удаленной автоматизированной обработки контента.

Сформулированные принципы стали определяющими при выборе программных средств разработки семантически структурированных сайтов. В качестве платформы проведенных разработок, была выбрана известная и популярная платформа Mediawiki версии 1.16.2. Платформа Mediawiki является легко расширяемой платформой для коллективной разработки текстовых документов любого вида, поддерживающая весь цикл разработки документа, включающий хранение различных версий документа, средства редактирования документов, удобные внутренние средства разметки документов для размещения в сети Internet. Для удаленной работы с документами платформа Mediawiki предоставляет сервис-ориентированное API, разработанное на базе REST-архитектуры. Основным языком разработки платформы является язык PHP, который в настоящее время является одним из основных средств разработки web-сайтов для сети Internet. В качестве платформы для хранения данных Mediawiki использует несколько различных современных СУБД, среди которых MySQL, PostgreSQL, SQLite. Выбранная платформа эффективно используется во многих Open source проектах, самым известным, из которых, является

проект Wikipedia, развиваемый сообществом Wikimedia Foundation.

Для апробации результатов, описанных в рамках данной работы, в выбранной платформе были разработаны программные средства, которые обеспечивают представления сложноструктурированной информации с помощью языка SCnML. Для обеспечения переноса содержимого web-сайта на другие CMS, которые поддерживают язык OWL, был разработан специальный транслятор, работающий через REST API Mediawiki и позволяющий получать содержимое web-сайта в виде текстов языка OWL. Указанные программные средства были опробованы при разработке интеллектуальной справочной системы по геометрии и сайте конференций OSTIS.

ЗАКЛЮЧЕНИЕ

В работе описан подход к построению web-сайтов на основе семантических сетей, которые характеризуются наличием сложноструктурированного контента. Результаты, описанные в работе, могут быть использованы при решении задач разработки прикладных интеллектуальных систем, в качестве вспомогательных средств коллективной разработки баз знаний, а также для повышения качества информационной поддержки традиционных web-сайтов.

БИБЛИОГРАФИЧЕСКИЙ СПИСОК

- [Микроформаты, 2012] Вики о микроформатах [Электронный ресурс]. – 2012. - Режим доступа: http://microformats.org/wiki/Main_Page-ru. – Дата доступа: 30.11.2012
- [Микроданные, 2012] Schema.org [Электронный ресурс]. – 2012. - Режим доступа: <http://schema.org/>. – Дата доступа: 30.11.2012
- [RDFa, 2012] Adida, Ben RDFa in XHTML: Syntax and Processing A collection of attributes and processing rules for extending XHTML to support RDF /Ben Adida [et al] //W3C Working Group Note 14 October 2008 [Электронный ресурс]. – 2012. - Режим доступа: <http://www.w3.org/TR/2008/REC-RDFa-syntax-20081014> – Дата доступа: 30.11.2012
- [SMW+, 2012] Semantic Enterprise Wiki Homepage [Электронный ресурс]. – 2012. - Режим доступа: http://www.smwplus.com/index.php/Semantic_MediaWiki_Plus – Дата доступа: 30.11.2012
- [SMW, 2012] Semantic Wiki Homepage [Электронный ресурс]. – 2012. - Режим доступа: <http://semantic-mediawiki.org/>
- [KnowWE, 2012] KnowWE – Knowledge Wiki Environment ! [Электронный ресурс]. – 2012.– Режим доступа: <http://sourceforge.net/projects/knowwe/> – Дата доступа: 30.11.2012
- [Колб, 2009] Колб, Д.Г Средства просмотра баз знаний интеллектуальных систем/ Д.Г. Колб//Вестник БрГТУ.–Минск, 2009. №5 (59).– С. 58-62
- [Белнап, 1981] Белнап, Н. Логика вопросов и ответов/ Н. Белнап, Т. Стил // Москва – “Прогресс”, 1981, 288 стр.
- [Сулейманов, 2001] Сулейманов, Дж.Ш. Исследование базовых принципов построения семантического интерпретатора вопросно-ответных текстов на естественном языке в АОС // Education Technology & Society 4(3). 2001.- с. 178-192
- [Anyanwu, 2007] Anyanwu K. SPARQL2: Towards Support for Subgraph Extraction Queries in RDF Databases./K. Anyanwu, A. Maduko, A. Sheth. // WWW 2007, P. 797–806.
- [Alkhateeb, 2009] Alkhateeb, F., Extending SPARQL with regular expression patterns (for querying RDF)/ F. Alkhateeb, J.-F.

Baget, J. Euzenat.// In Web Semantics: science, Services and Agents on the World Wide Web. – 2009. – № 7(2). – P 57–73

[Pérez, 2010] Pérez, Jorge nSPARQL: A Navigational Language for RDF/ Jorge Pérez, Marcelo Arenas, Claudio Gutierrez// Journal of Web Semantics. – 2010, 8(4). – P. 255-270.

[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.

[Голенков, 2011] Голенков, В. В. Принципы построения массовой семантической технологии компонентного проектирования интеллектуальных систем / В.В. Голенков, Н. А. Гулякина //Материалы международной научно-технической конференции OSTIS-2011, Мн.: БГУИР, 2011. - 21-58 стр.

[Трофимов, 2011] Трофимов, И. В. Эволюция выразительных способностей языка OWL /И. В. Трофимов // Программные системы: теория и приложения : электрон. научн. журн.2011. № 4(8), с. 85–94. URL: http://psta.psriras.ru/read/psta2011_4_85-94.pdf

DEVELOPMENT OF WEB-SITES WITH A COMPLEX CONTENT BASED ON SEMANTIC NETWORKS

Kolb D.G., Furman O.D.

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

kolb@bsuir.by

olgasharpio@gmail.com

The approach to development of semantic structured web sites are offered. The model of web sites of this class is briefly described, the idea the organization of navigation search in web sites of this class is considered. The method for mapping of a class of the semantic networks used for development of semantic structured web sites, to the RDF model is offered.

INTRODUCTION

The problem of development of semantic web sites even more often enters into the sphere of interests of many scientists of the whole world. On the one hand the solution of this problem will allow to pass to new qualitative level of the World Wide Web – Semantic Web, on the other hand will allow to overcome one more step on a way to creation of intellectual systems and to put a platform for creation of intellectual systems for Internet.

Life cycle of process of development of web sites includes the following main objectives:

- filling of a content of a web site (including multimedia);
- optimization of a content for search engines;
- development of tools of navigation and search in a web site;
- providing shipping of a content of a web site.

Modern WCMS solve the specified problems with various level of completeness, however there is also a number of not solved tasks which decision can't be received without use of methods and artificial intelligence techniques.

MAIN PART

Considering solutions which are proposed now for improvement of web sites with a complex content, it should be noted two significant directions of works:

1. development of methods and tools of the specification of pages of a web site which allow to carry out easy indexation of web pages and, respectively, to provide more relevant search;

2. development of methods and tools of the specification of information which allow to represent complex types of information, including, such as theorems, axioms, logical statements and etc.

There is a dilemma – to use expressionless tools which are adapted for problems of industrial development of web sites or to use tools possessing the bigger expressive capacity, however not adapted for industrial development.

Within this work authors will try to propose the solution which will allow to integrate two directions of development of web sites given above, on the basis of use of models of representation of knowledge of higher abstract level. Such approach from the point of view of authors will allow to receive with one the intermediate decision for applied industrial development of web sites on the basis of semantic networks. Also will give an impetus to an evolution of development tools of the web sites based on semantic networks existing now.

Semantic structured web sites we will call a special class of the intellectual systems constructed on the basis of semantic networks and hypertext model of representation of information. The model of semantic structured web site is set as follows:

$$M = \{O, S_A, S_P, O_{hyp}, M_L, M_S, I\}, \quad (1)$$

where

O – ontology of web site's subject area;

S_A – a set of classes of articles of semantic structured web site (atomic semantic units);

S_P – a set of classes of pages of semantic structured web site, each of which consists of some set of classes of articles and each element which is connected with other element the sequence relation;

O_{hyp} – ontology of multimedia documents;

M_L – language tools of submission of information in the form of uniform semantic networks;

M_S – model of semantic navigation search on semantic structured web site;

I – method for mapping of model of semantic structured web site to the RDF model.

CONCLUSION

In work approach to creation of web sites on the basis of semantic networks which are characterized by existence of a complex content is described. Results described in article can be used at the decision, problems of development of applied intellectual systems, as supportive applications of collective development of knowledge bases, and also for improvement of quality of information support of traditional web sites.