ISSN 1814-1196
Научный вестник НГТУ
том 72, № 3, 2018, с. 43–58

http://journals.nstu.ru/vestnik
Science Bulletin of the NSTU
Vol. 72, No. 3, 2018, pp. 43–58

ИНФОРМАТИКА,
ВЫЧИСЛИТЕЛЬНАЯ ТЕХНИКА
И УПРАВЛЕНИЕ

INFORMATICS,
COMPPUTER ENGINEERING
AND CONTROL

УДК 004.7.056.53

DOI: 10.17212/1814-1196-2018-3-43-58

Метод онтологического анализа web-ресурса на основе метаданных*

В.И. ВОРОБЬЕВ[1, a], А.А. СОЛДАТКИНА[2, b]

[1] *199178, РФ, Россия, г. Санкт-Петербург, 14-я линия, 39, Санкт-Петербургский институт информатики и автоматизации Российской академии наук*
[2] *197376, Россия, Санкт-Петербург, ул. профессора Попова, 5, Санкт-Петербургский государственный электротехнический университет им. Ленина*

[a] *vvi@iias.spb.su* [b] *alinasoldatkina2014@gmail.com*

Важной проблемой современного интернет-сообщества является распространение огромного количества информации, что вызывает трудности для быстрого поиска достоверных знаний. В данной работе предложен новый метод технологии анализа и обработки данных, который, основываясь на семантических связях, ускоряет вывод необходимой информации, а также оценивает надежность ее источников. Особое внимание в статье уделено рассмотрению применяемых на сегодняшний день методов анализа web-контента. В статье предлагается использовать новый метод, который основывается на выделении метаинформации с web-сайта и рассмотрении ее семантических связей. Для этого в редакторе Protege 5.0 разработана семантическая модель, содержащая большое количество классов и свойств, характерных для элементов данной предметной области. В работе рассмотрены все основные этапы построения онтологической модели предметной области, выделены методы анализа и классификации web-ресурсов, приведены примеры описания классов и содержащихся в них экземпляров, отношений между ними. Для автоматической классификации разработаны логические правила, которые проверяют семантические связи между метаданными ресурса и наборами ключевых слов классов. Надежность источника определяется исходя из набора и объема его метаданных, что позволяет оценить достоверность и качество представленного контента. Предложенный онтологический подход является перспективным с точки зрения высокого уровня интероперабельности информационных систем за счет открытых интерфейсов доступа, а также путем использования единого формата записи и обмена данными. В рамках онтологического подхода семантическая способность к взаимодействию реализована на основе единого представления информации в предметной области. Для повышения скорости и точности вывода поисковых запросов предлагается использовать запросы из семантической базы данных на языке SPARSQL, примеры которых также приводятся в статье.

Ключевые слова: семантические технологии, онтология, rdf, owl, метаданные, уровень доверия, SPARSQL, анализ web-ресурсов, квалиметрическая шкала

* *Статья получена 27 мая 2018 г.*

ВВЕДЕНИЕ

В эпоху формирования цифровой инфраструктуры информационного общества информация становится ключевым объектом, необходимым для успешного развития государства и общественной жизни. Объемы информации, циркулирующие в web-пространстве, динамично развиваются, и ее объем удваивается ежегодно. При этом возрастает доля ложной, недостоверной и заведомо искаженной (фейковой) информации [1, 2]. Важно то, что объемы подобной информации достигли критических значений – это открывает возможность манипулирования информацией и ее восприятием.

Фактически 49 % пользователей социальных сетей в США в 2012 году получили ложные новости через социальные сети. Аналогичным образом согласно опросу Silverman 11, проведенному в 2015 году, ложные слухи и дезинформация распространяются быстрее, чем когда-либо, из-за социальных сетей. Политические аналитики продолжают обсуждать дезинформацию и поддельные новости в социальных сетях и их влияние на президентские выборы в США в 2016 году [3]. Международная перепись населения с 2017 года насчитывала 114 активных служб проверки фактов, что на 19 % больше по сравнению с предыдущим годом.

Важной задачей современного информационного общества является развитие и разработка методов, которые бы позволили быстро анализировать ресурсы, оценивать надежность и качество представленных в них данных.

Актуальность задачи подтверждается наличием ряда проектов разных стран. В частности, в рамках европейской программы HORIZON 2020 разрабатывается проект AEGIS, направленный на создание взаимосвязанной цепочки ценностей данных, в котором создается новая платформа для большого объема данных, интеграции, анализа и обмена данными на разных языках, очистки данных, связывания семантических данных, построения системы ценности данных пользователей и организаций и внедрения новых бизнес-моделей для экономики обмена данными. Проект AEGIS стремится повысить ценность данных, хранящихся на его платформе, путем семантического обогащения их более полезной информацией, и это будет сделано с использованием хорошо установленных стандартов и технологий.

Качественный анализ содержимого web-ресурсов занимает большое количество времени и требует существенных затрат людских и материальных ресурсов. Для автоматической обработки web-документа можно внедрить подход, основанный на семантическом анализе метаданных [4], которые занимают малый процент от общего количества контента, но тем не менее содержат описания основных и дополнительных свойств web-ресурса (его происхождение, назначение, связи, имиджевый ресурс, промосайт, лендинг (landing page), корпоративный веб-сайт, информационный портал, интернет-магазин, «business-to-business»-представительство, «виртуальный офис»).

Метаданные рассредоточены по всему web-документу, поэтому важной задачей является описать их, не потеряв никаких семантических связей [5]. Для решения данной задачи применяются семантические технологии. Мета-информация, представленная в виде семантической сети, ускорит поиск необходимой информации, сделает ее более безопасной и надежной, повысит релевантность выводимых данных.

1. МЕТОДЫ АНАЛИЗА WEB-РЕСУРСОВ

Важнейшим элементов любой web-страницы является ее контент, т. е. наполнение. В контент входят текст, графика, видео-, аудио- и другая информация. Качественное наполнение сайта очень сильно влияет на его рейтинги и отзывы пользователей [6]. Для оценки контента используют множество средств и методов: анализ юзабилити сайта, анализ семантического ядра запросов, анализ сайта к индексации и видимости машинами, анализ посещаемости сайта и сбор статистики.

Специалисты области анализа сайтов выделяют два основных метода – количественный и качественный.

Количественный анализ позволяет оценить сайт в «цифрах». Он показывает, сколько структурных единиц было задействовано в создании сайта. Оценивается количество слов, плотность и размер текста, наличие ключевых слов и многое другое.

Качественным методом можно оценить контент сайта с точки зрения пользователя. Обращается внимание на уровень изложения материала, его точность и уникальность, оцениваются легкость подачи информации и актуальность. Важным показателем является соответствие содержания текста заголовку [7]. Оценка качественных показателей вызывает большие трудности.

Существует несколько методов проверок соответствия страниц поисковым запросам пользователей – «от поисковиков» и «от сайта». Первый метод не анализирует сайт, который еще не был проиндексирован поисковой системой, да и сама поисковая система не может интерпретировать так же качественно, как человек. Конечно, сейчас активно ведутся работы в области искусственного интеллекта и нейронных сетей, но на практике внедрить их непросто. В базах данных информация содержится в структурированном виде, но при передаче ее на интернет-страницу многие связи теряются и значительная часть важной информации растворяется в общем потоке текста. Для неструктурированной информации поисковики придают больший вес ключевым словам сайта даже при частичном использовании семантического анализа.

Второй метод проверки (качественный) оценивает уровень слов и словосочетаний, отражающих структуру и тематику сайтов. Он идеален для только созданных web-страниц и берет во внимание интересы пользователей. Но такой метод очень трудоемок, и автоматизировать его очень сложно.

На сегодняшний день существует много онлайн-сервисов по проверке репутации сайта. Задачу определения уровня доверия решает сам пользователь, и, учитывая его недостаточную осведомленность, доверие формируется случайным набором факторов (интуицией, навязанным мнением). Особенно ярко это проявляется в социальных сетях, блогах, новостных сводках. Поэтому проблема доверия к предоставляемой информации остро стоит перед всеми пользователями. По данным Всероссийского центра общественного мнения за 2017 год, лишь 19 % опрошенных доверяют новостным, аналитическим и официальным сайтам, 22 % не доверяют, остальные затрудняются в однозначном ответе.

С доверием связаны вопросы и идентификации, и безопасной передачи информации, и контроля доступа, поэтому люди склонны подвергать сомнению информацию из непроверенных источников, которыми являются в большинстве своем web-ресурсы. Так как штатные средства не обеспечивают

должного уровня безопасности, то пользователь прибегает к своему личному опыту оценки доверия.

Грамотное использование метаданных позволяет повысить уровень доверия, так как содержит структурную информацию о контенте.

2. ПРИМЕНЕНИЕ СЕМАНТИЧЕСКИХ ТЕХНОЛОГИЙ ДЛЯ ОПИСАНИЯ И АНАЛИЗА МЕТАДАННЫХ

Для описания большого количества неструктурированных данных применяются технологии Semantic Web, содержащие инструменты для представления данных и связей между ними (рис. 1) [8].

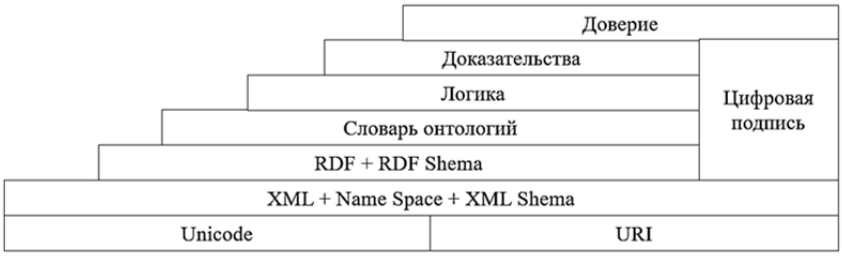

Рис. 1. Архитектура Semantic Web

Fig.1. The architecture of Semantic Web

Представленная архитектура обозначает инструментальный фундамент, на котором строится доверие к web-ресурсу. В возможности семантического веба входят семантический поиск, объединение данных, логический вывод [9].

На языках XML и RDF данные записываются в виде триплетов, но их возможностей недостаточно, чтобы описать всю структуру и все связи web-данных. Язык OWL позволяет создавать онтологии в терминах классов и свойств, поддерживает описание простых логических проверок целостности онтологий предметной области и их связей друг с другом, а также импорт внешних определений.

Онтологии явно описывают метаданные web-ресурсов и их отношения. По своей структуре онтология представляет собой набор некоторых категорий: классы, отношения, функции, аксиомы, экземпляры. Онтологический анализ строится на разделении данных на классы с последующим выделением подклассов и экземпляров данных классов.

Важным элементом онтологий являются отношения, функции и аксиомы. Одним из самых распространенным видов связей являются отношения категоризации, имеющие несколько названий: отношение is-a, класс – подкласс, родовое отношение, отношение a-kind-of [10]. Функции – это вид отношений, с помощью которых можно выразить, как экземпляры связаны между собой, в каких ролях они находятся. Аксиомы нужны для выражения утверждений, которые связывают понятия и свойства. Они позволяют продемонстрировать информацию, которую очень трудно описать другими методами, такими как построение иерархий понятий или установки отношений и функций.

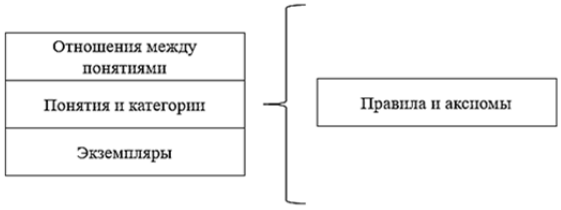

Рис. 2. Структура онтологий

Fig. 2. The structure of ontologies

Экземпляры в онтологии выступают в роли индивидуумов, под которыми подразумеваются конкретные элементы класса. Все элементы подчиняются иерархии, представленной на рис. 2.

3. ВЫДЕЛЕНИЕ МЕТАДАННЫХ ИЗ WEB-РЕСУРСА

Контент ресурса (текст, аудио, изображение и др.) имеет метаданные, которые не всегда видны пользователю, но которые необходимы для поиска и структуризации информации [11, 12]. Так, например, метаданные позволяют построить фолксономию сайта, т. е. «категорировать» данные за счет произвольно выбираемых тегов.

Все метаданные, содержащиеся в web-ресурсе, можно разделить на четыре категории (рис. 2).

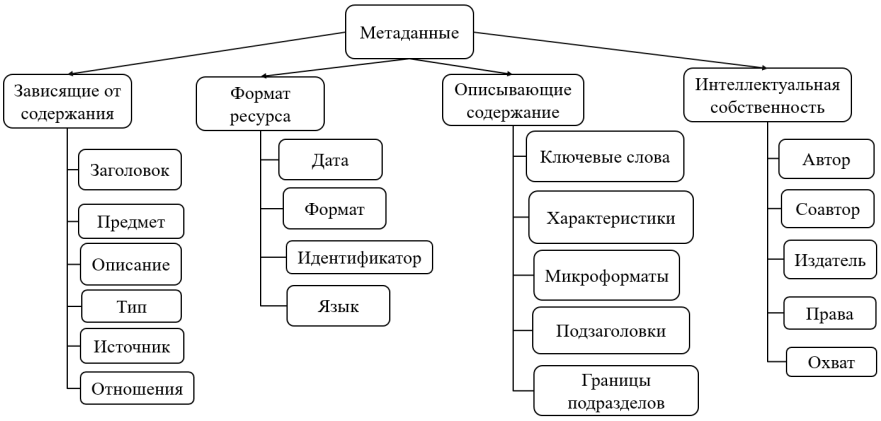

Рис. 3. Метаданные web-ресурсов

Fig. 3. Web-resource metadata

Стандартом описания метаданных web-документов является так называемое дублинское ядро, содержащее пятнадцать элементов. Особенностью такого описания является то, что элементы могут повторяться или не содержаться вовсе.

Пример описания web-страницы с использованием элементов дублинского ядра:

```
<META NAM E="DC. Title" CONTENT="Каталог образовательных
ресурсов сети Интернет">
```

```
<META NAME="DC. Subject" CONTENT="Образовательные ресурсы сети Интернет">
<META NAME="DC. Keywords" CONTENT="федеральные образовательные ресурсы, ресурсы по предметам образовательной программы ">
<META NAME="DC. Description" CONTENT="Каталог образовательных ресурсов сети Интернет для основного общего и среднего (полного) общего образования. ">
<META NAME="DC.Creator " CONTENT=" EDU-TOP">
<META NAME="DC. Publisher. CorporateName" CONTENT="Первый независимый общероссийский рейтинг-каталог школьных сайтов EDU-TOP.ru">
<META NAME="DC. Contributor. PersonalName" CONTENT="Иванов. А. Н ">
<META NAME="DC. Date. Query" CONTENT="2018-04-09">
<META NAME="DC. Format" CONTENT="text/html">
<META NAME="DC. Format" CONTENT="23 976  bytes">
<META NAME="DC. Type" CONTENT="Text">
<META NAME="DC. Type" CONTENT="Изображение X-ICON">
<METANAME="DC.Source"CONTENT="https://yandex.ru/clck/jsredir?bu=uniq15227517223851266254 3&from=yandex.ru%3Bsearch%2F%3Bweb%3B%3B&text=&etext=1746.C0">
<META NAME="DC. Identifier" CONTENT="http://edu-top.ru/katalog/">
<META NAME="DC. Language"  CONTENT="ru">
<META NAME="DC. Relation" http://edu-top.ru/katalog/favicon.ico">
<META NAME="DC. Coverage. PlaceName" Российская Федерация">
```

Микроформаты тоже являются метаданными ресурса, так как это один из способов семантической разметки сведений о разных объектах (событиях, организациях, людях, товарах и т. д.) на web-ресурсах с помощью стандартных элементов языка html и xhtml [8]. Интеллектуальные агенты способны извлечь из таких форматов структурированную информацию, следуя определенным правилам и соглашениям. Поскольку микроформаты основаны на уже существующих стандартах, их легко добавлять на существующие страницы web-ресурсов.

4. ПОСТРОЕНИЕ СЕМАНТИЧЕСКОЙ МОДЕЛИ

Для семантического описания метаданных и написания онтологий удобно использовать редактор Protege.

Этапы построения онтологий: создание классов и определение подклассов, создание свойств, занесение экземпляров, определение форматов, определение связей между экземплярами и классами, создание аксиом, формирование запросов [13].

На первом этапе нужно определить основные классы, а также выделить подклассы. На основе таксономических отношений строится дерево классификации понятий, иначе говоря, иерархия классов. Корень дерева – web-ресурс, его подклассы – классификация web-ресурсов и метаданные (рис. 4).

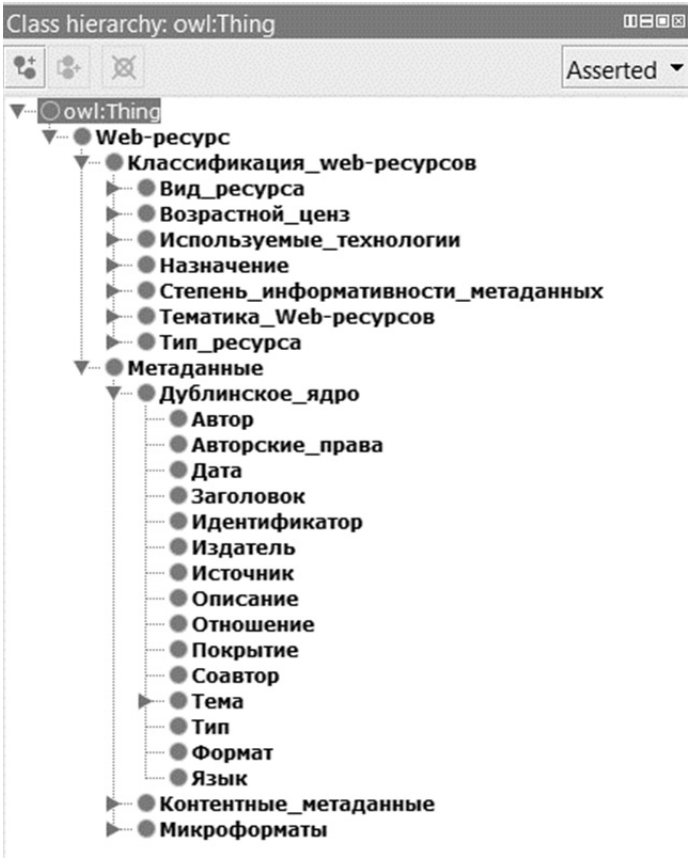

Рис. 4. Классы онтологической модели

Fig. 4. Ontological model classes

Редактор Protege позволяет наглядно увидеть классы и связи между ними с помощью механизмов построения семантической сети. Семантическая сеть представляет собой сетевую модель предметной области (ПрО), которая имеет вид ориентированного графа, узлами которого являются классы, а ребра – связями (рис. 5).

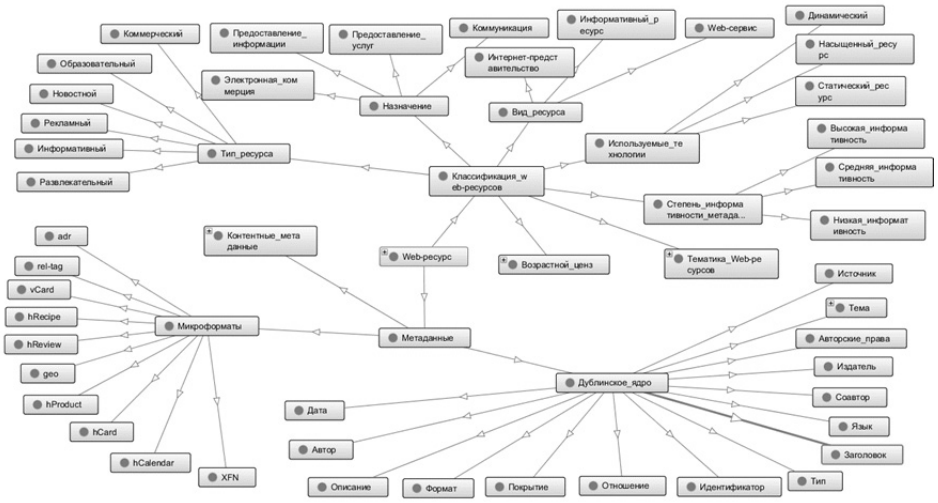

Рис. 5. Пример иерархии классов web-ресурсов

Fig. 5. An example of hierarchy of Web-resource classes

В онтологических моделях классы представляют собой общие понятия о ПрО. Каждый класс описывает индивидуальность сущностей, которые имеют общие атрибуты или свойства, например, web-ресурс и его автор. На вкладке Object Property создаются онтологии (рис. 6).

Рис. 6. Панель описания множества свойств, присущих объектам ПрО

Fig. 6. A panel of description of a set of properties characteristic of objects

На языке rdf это выглядит следующим образом:

```
<SubObjectPropertyOf>
    <ObjectProperty IRI="#имеет_автора"/>
    <ObjectProperty IRI="#Дублинское_ядро"/>
</SubObjectPropertyOf>
```

При создании свойства определяются его диапазон и домен из существующих классов, а также дополнительные характеристики. Например, свойству «имеет_автора» соответствует онтология – противоположное свойство «автор_ресурса».

С помощью OntoGraf можно наглядно увидеть связь свойств и классов. На рис. 7 представлены связи между web-ресурсом и классами классификаций, а на рис. 8 – между web-ресурсом и элементами дублинского ядра.

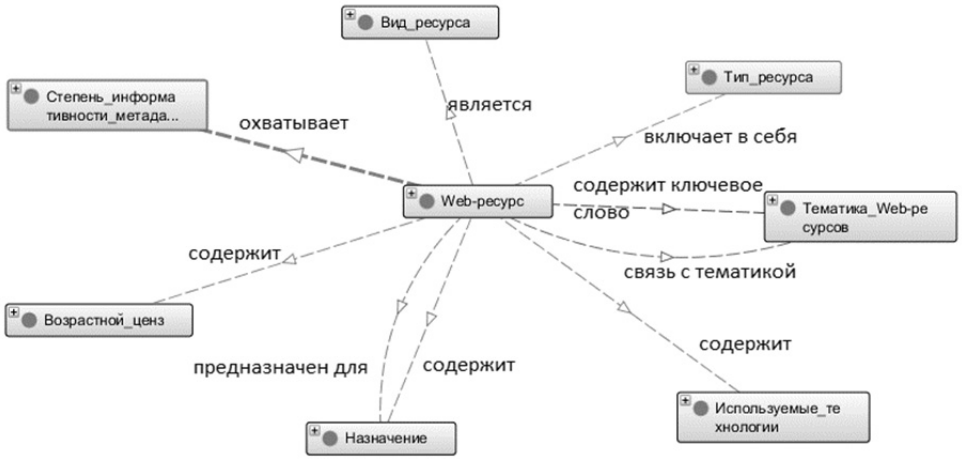

Рис. 7. Фрагмент визуализации онтологии классификации web-ресурса

Fig. 7. A visualization fragment of the Web-resource classification ontology

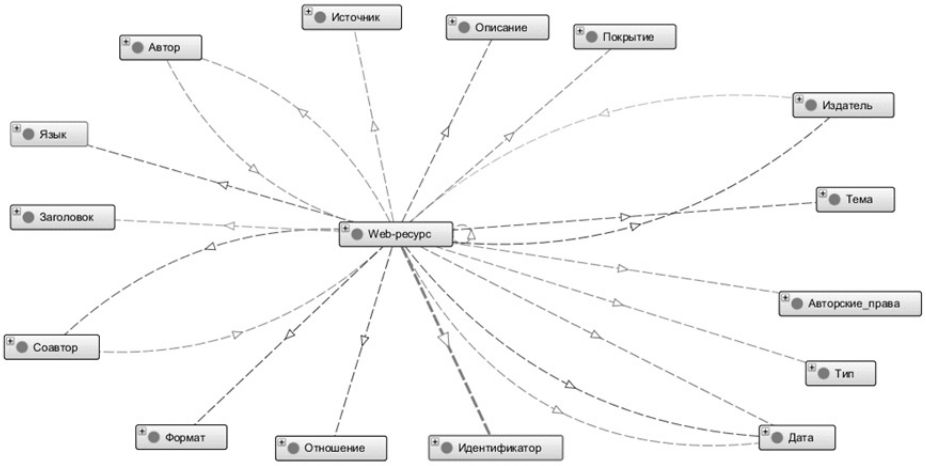

Рис. 8. Фрагмент визуализации онтологии описания web-ресурса

Fig. 8. A visualization fragment of the Web-resource description ontology

Индивиды в редакторе Protege определяются с помощью аксиом фактов: факты членства в классах, факты о значении свойств индивидов.

Пример аксиом первого вида:

```
<ObjectPropertyAssertion>
    <ObjectProperty IRI="#имеет_язык"/>
    <NamedIndividual IRI="#102058_Web-ресурс"/>
    <NamedIndividual IRI="#ru"/>
</ObjectPropertyAssertion>
```

Аксиомы устанавливают, к каким классам принадлежит объект и какими свойствами он обладает, а также определяют связи между объектами (рис. 9).

Рис. 9. Окно описания экземпляров

Fig. 9. A window of copy description

Значения экземпляров задаются с помощью новых свойств, таких как owl:DatatypeProperty. Для этого используются аксиомы второго вида:

```
<DataPropertyAssertion>
    <DataProperty IRI="#телефон"/>
    <NamedIndividual IRI="#Смирнов_Сергей_Борисович"
/>                                          <Literal
datatypeIRI="http://www.w3.org/2001/XMLSchema#string">+
7(495)984 81 36
</Literal>
    </DataPropertyAssertion>
```

Для создания и определения значений нужно перейти на вкладку Data Properties (рис. 10).

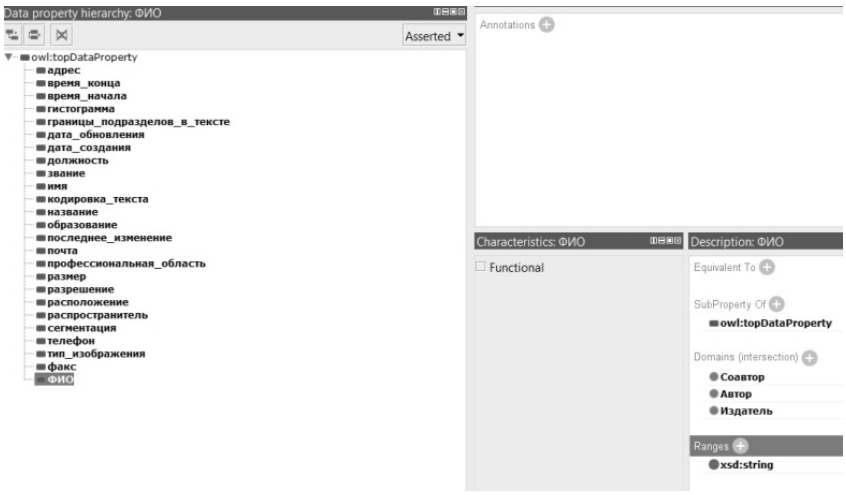

Рис. 10. Окно значений экземпляров

Fig. 10. A window of copy values

Возможности логического вывода редактора Protege «по умолчанию» довольно скромны, поэтому для вывода неявных следствий онтологии применяют SWRL-правила. Каждый класс имеет некоторые ключевые слова, присущие именно ему и, если ресурс содержит эти ключевые слова, то мы относим его к классу. Такие логические правила позволяют создавать новые триплеты.

Пример SWRL-правила:

```
содержит_ключевое_слово(?x, ?y) ^ Тематика_Web-ресурсов(?y) ->Тематика_Web-ресурсов(?x)
```

Данное правило проверяет, содержит ли ресурс ключевые слова, которые принадлежат определенной теме, и если содержит, то ресурс также можно отнести к этой теме.

Редактор Protege позволяет в интерактивном режиме вводить sql запросы. Для этого нужно перейти на вкладку SPARQL Query. Язык SPARQL очень похож на стандартный язык запросов к БД SQL, но имеет и несколько существенных отличий. SPARQL – язык запросов к RDF хранилищам, т. е. к данным, представленным в виде RDF-триплетов. До начала работы следует прописать PREFIX, который будет служить для указания сокращений универсальных идентификаторов ресурса (URI).

Редактор Protege позволяет работать с четырьмя типами запросов к данным: запрос «по умолчанию», запрос с применением информации о литерале, запрос на основе данных об индивидууме, запрос на основе данных об объектных свойствах классов.

Примеры

Показать все элементы дублинского ядра для одного web-ресурса

Запрос:

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
```

```
PREFIX
:<http://www.semanticweb.org/user/ontologies/2018/4/ont
ology-web#>
SELECT distinct ?метаданные ?значения
WHERE { :103679_Math-Net.Ru ?метаданные ?значения .
        ?метаданные rdf:type owl:ObjectProperty }
```

метаданные	значения
имеет_язык	ru
имеет_покрытие	Москва,Россия
имеет_соавтора	Смирнов_Сергей_Борисович
имеет_формат	text/html;_charset=UTF-8
авторские_права_принадлежат	Отдел_компьютерных_сетей_и_информационных_технологий
имеет_издателя	Математический_институт_им._В.А._Стеклова_Российской_академии_наук
имеет_ключевое_слово	Математический_институт_им._В._А._Стеклова_РАН
имеет_описание	Общероссийский_математический_портал_Math-Net.Ru_—_это_современная_информационная_систе
имеет_дату_создания	2003-09
имеет_заголовок	Общероссийский_математический_портал
имеет_тип	Text
имеет_тему	Общероссийский_математический_портал
имеет_заголовок	Math-Net.Ru
имеет_идентификатор	URL:http://www.mathnet.ru/
имеет_автора	Чебуков_Дмитрий_Евгеньевич

Рис. 11. Ответ на запрос по выводу дублинского ядра

Fig. 11. Answer to the inquiry about the Dublin core output

Показать проклассифицированный web-ресурс
Запрос:

```
SELECT distinct ?класс ?значение
WHERE {:102058_Web-ресурс rdf:type ?значение .
?значение rdfs:subClassOf ?класс}
```

класс	значение
Назначение	Предоставление_информации
Тип_ресурса	Развлекательный
Используемые_технологии	Насыщенный_ресурс
Степень_информативности_метаданных	Средняя_информативность
Возрастной_ценз	0+

Рис. 12. Вывод классификаций одного ресурса

Fig. 12. Output of one resource classifications

Вывести всю информацию об авторе определенного web-ресурса
Запрос:

```
SELECT distinct ?автор ?информация ?значение
WHERE {:103679_Math-Net.Ru :имеет_автора ?автор .
    ?автор ?информация ?значение.
    ?информация rdf:type owl:DatatypeProperty }
```

автор	информация	значение
Чебуков_Дмитрий_Евгеньевич	должность	"зав. информационно-издательским сектором"^^<http://www.w3.org/2001/XMLSchema#string>
Чебуков_Дмитрий_Евгеньевич	звание	"кандидат хим. наук"^^<http://www.w3.org/2001/XMLSchema#string>
Чебуков_Дмитрий_Евгеньевич	телефон	"+7 (495) 984 81 41"^^<http://www.w3.org/2001/XMLSchema#string>
Чебуков_Дмитрий_Евгеньевич	почта	"tche@mi.ras.ru"^^<http://www.w3.org/2001/XMLSchema#string>

Рис. 13. Вывод информации об авторе

Fig. 13. Output of information about the author

Вывести возрастной ценз ресурсов
Запрос:

```
SELECT ?ресурс ?class
WHERE {?ресурс rdf:type :Web-ресурс.
    ?ресурс rdf:type ?class.
    ?class rdfs:subClassOf :Возрастной_ценз}
```

ресурс	class
102058_Web-ресурс	0+
103679_Math-Net.Ru	6+

Рис. 14. Вывод возрастного ценза ресурса

Fig. 14. Output of the resource voting age

Вывести ресурс, издатель которого имеет определенный номер телефона
Запрос:

```
SELECT ?ресурс ?издатель ?информация
WHERE {?ресурс rdf:type :Web-ресурс.
    ?ресурс :имеет_издателя ?издатель.
    ?издатель :телефон ?информация.
FILTER(str (?информация) = "89872345634")}
```

ресурс	издатель	информация
103679_Math-Net.Ru	Математический_институт_им._В.А._Стеклова_Российской_академии_наук	"89872345634"^^<http://www.w3.org/2001/XMLSchema#integer>

Рис. 15. Вывод определенного ресурса

Fig. 15. Output of a certain resource

Задачу количественной оценки доверия (рисков использования web-ресурса) можно построить с помощью введения квалиметрических шкал. Как показано выше, мы можем построить наборы тестовых вопросов для определения уровня доверия каждого используемого информационного элемента. Далее достоверность можно оценивать с помощью квалиметрической шкалы. В качестве примера рассмотрим случай: когда данные из всех источников совпадают, отсчет на шкале будет иметь значение «1», а если все источники дают разные данные, то значение «0». Промежуточные значения в зависимости от совпадения могут принимать значения от нуля до единицы [14, 15].

В результате мы имеем возможность строить вектор показателей уровня доверия сложного ПрО в виде $q = (q1,\ldots, qn)$, $qi\ (x)i = 1,\ldots,l$ вектора исходных характеристик $x = (x1,\ldots, xn)$, где xi – i-минимальный информационный элемент web-ресурса. Например, для оценки достоверности данных сайта применяются набор показателей q в зависимости от x (характеристики: происхождение, назначение, связи и т. д.). После получения набора отдельных показателей выбирается синтезирующая функция:

$$Q\ (q) = Q\ (q; w),$$

где $w = (w1,\ldots, wl)$, $w1,+ \ldots + wl = 1$ рассматриваются как весовые коэффициенты, задающие степень влияния отдельных показателей на сводную оценку.

В случае дефицита информации при неопределенности выбора функций q, Q и вектора w квалиметрическая шкала имеет более бедную структу-

ру, чем обычная числовая шкала (например, показатели качества сайта в зависимости от его места его размещения). В этом случае задача оцифровки состоит в выборе отображения φ (b), где b – качественная характеристика (например, баллы).

ЗАКЛЮЧЕНИЕ

Разработанная прикладная онтология позволяет решать задачи логического и структурированного описания информации, а также категоризации, быстрого поиска и анализа данных. С помощью языка OWL спроектирована информационная система, оперирующая знаниями, содержащимися в онтологиях. Показана возможность построения базы знаний ПрО. Приведены примеры запросов, которые позволяют нам определить уровень доверия. Приводится алгоритм для его количественного определения. В будущем онтология может использоваться в качестве баз знаний для формирования единой среды интеллектуальной системы для классификации web-данных. Разработана классификация web-данных. Использование интеллектуальных агентов позволит в будущем формулировать запрос и оценивать ответ на него с помощью семантических связей и ограничений, описанных в онтологии.

СПИСОК ЛИТЕРАТУРЫ

1. *Осипов В.Ю., Воробьев В.И., Левоневский Д.К.* Проблемы защиты от ложной информации в компьютерных сетях // Труды СПИИРАН. – 2017. – Вып. 4 (53). – С. 97–117.

2. *Бергель Х.* Брюс Шнаер о цифровых угрозах будущего // Открытые системы. СУБД. – 2018. – № 1. – С. 34.

3. *Brandtzaeg P., Følstad A.* Trust and distrust in online fact-checking services // Communications of the ACM. – 2017. – Vol. 60, N 9. – P. 65–71.

4. *Тузовский А.Ф.* Архитектура семантического web-портала // Известия Томского политехнического университета. – 2006. – Т. 309, № 7. – С. 142–145.

5. *Gruber T.* Collective knowledge systems: where the Social Web meets the Semantic Web // Journal of Web Semantics. – 2008. – Vol. 6, N 1. – P. 4–13.

6. *Ландэ Д.В.* Поиск знаний в Internet. – М.: Диалектика, 2005. – 272 с. – (Профессиональная работа). – ISBN 5-8459-0764-0.

7. *Рогушина Ю.В., Гришанова А.Ю.* Средства интеллектуализации поиска информационных ресурсов в сети Интернет // Сборник трудов VII Международной конференции «Интеллектуальный анализ информации ИАИ-2007». – Киев, 2007. – С. 322–331.

8. *Yi L., Liu B.* Web page cleaning for web mining through feature weighting // Proceedings of Eighteenth International Joint Conference on Artificial Intelligence (IJCAI-03). – Acapulco, Mexico, 2003. – P. 43–48.

9. *Gladun A., Rogushina J., Shtonda V.* Ontological approach to domain knowledge representation for information retrieval in multiagent systems // Information Theories and Applications. – 2006. – Vol. 13, N 4. – P. 354–362.

10. *Gruber T.* Toward principles for the design of ontologies used for knowledge sharing? // International Journal Human-Computer Studies. – 1995. – Vol. 43. – P. 907–928.

11. *Гото К.* Веб-редизайн: книга Келли Гото и Эмили Котлер. – СПб.: Символ-Плюс, 2003. – 376 с.

12. Семантика, метаданные и онтологии в приложениях для умного города – новые стандарты BSI / В.П. Куприяновский, Д.И. Ярцев, А.А. Харитонов, Н.А. Уткин, Д.Е. Николаев, В.И. Дрожжинов, Д.Е. Намиот, Ю.И. Волокитин // International Journal of Open Information Technologies. – 2017. – Т. 5, № 6. – С. 94–108.

13. *Горшков С.С.* Введение в онтологическое моделирование. – [Б. м.]: ООО ТриниДата, 2016. – 165 с.

14. *Хованов Н.В.* Анализ и синтез показателей при информационном дефиците. – СПб.: Изд-во С.-Петерб. ун-та, 1996. – 196 с.

15. *Hovanov N., Yudaeva M., Hovanov K.* Multicriteria estimation of probabilities on basis of expert non-numeric, non-exact and non-complete knowledge // European Journal of Operational Research. – 2009. – Vol. 195, iss. 3. – P. 857–863.

Воробьев Владимир Иванович, профессор, главный научный сотрудник Санкт-Петербургского института информатики и автоматизации Российской академии наук. В списке научных трудов более 115 работ в области математического моделирования и информатики. E-mail: vvi@iias.spb.su

Солдаткина Алина Андреевна, студент магистратуры Санкт-Петербургского государственного электротехнического университета. E-mail: alinasoldakna2014@ gmail.com

DOI: 10.17212/1814-1196-2018-3-43-58

Method of ontological analysis of a web-resource based on metadata[*]

V.I. VOROBYEV[1, a], *A.A. SOLDATKINA*[2, b]

[1] *St. Petersburg, St.-Petersburg Institute of Informatics and Automation of the Russian Academy of Sciences, 14 Line, 39, St. Petersburg, 199178, Russia*
[2] *St.-Petersburg State Electrotechnical University named after Lenin, Professor Popov St., Building 5, St. Petersburg, 197376, Russia*

[a] *vvi@iias.spb.su* [b] *alinasoldatkina2014@gmail.com*

Abstract

An important problem of the modern Internet community is the dissemination of a huge amount of information, which makes it difficult to quickly find the right knowledge. In this paper we propose a new method of analysis and data processing technology that, based on semantic links, accelerates the output of necessary information, and also assesses the reliability of its sources. Particular attention in the article is given to the methods of analysis of web content that are used today. The article proposes using a new method which is based on the allocation of meta-information from the web-site and consideration of its semantic links. To do this, the Protégé 5.0 editor developed a semantic model that contains a large number of classes and properties characteristic of the elements of the given subject area. The paper considers all the basic stages of constructing an ontological model of the subject area, identifies methods for analyzing and classifying web resources, gives examples of the description of classes and the instances contained in them as well as the relationships between them. For automatic classification, logical rules are developed that check semantic links between the resource metadata and the sets of class keywords. The dependability of the source is determined based on its metadata set and volume, which allows you to evaluate the reliability and quality of the presented content. The proposed ontological approach is promising from the point of view of a high level of interoperability of information systems due to open access interfaces as well as by using a single recording format and exchange of data. Within the framework of the ontological approach, the semantic ability to interact is realized on the basis of a unified representation of information in the subject domain. To increase the speed and accuracy of the output of search queries, it issuggested using queries from the semantic database in the SPARSQL language, whose examples are also given in the article.

Keywords: semantic technologies, ontology, rdf, owl, metadata, confidence level, SPARSQL, web resources analysis, qualimetric scale

[*] *Received 27 May 2018.*

REFERENCES

1. Osipov V.Yu., Vorob'ev V.I., Levonevskii D.K. Problemy zashchity ot lozhnoi informatsii v komp'yuternykh setyakh [Problems of protection from false information in computer networks]. *Trudy SPIIRAN – SPIIRAS Proceedings*, 2017, iss. 4 (53), pp. 97–117.

2. Berghel H. Bryus Shnaer o tsifrovykh ugrozakh budushchego [Bruce Shnaer on future digital threats]. *Otkrytye sistemy. SUBD – Open Systems. DBMS*, 2018, no. 1, p. 34.

3. Brandtzaeg P., Følstad A. Trust and distrust in online fact-checking services. *Communications of the ACM*, 2017, vol. 60, no. 9, pp. 65–71.

4. Tuzovsky A.F. Arkhitektura semanticheskogo Webportala [The architecture of the semantic Web portal]. *Izvestiya Tomskogo politekhnicheskogo universiteta – Bulletin of the Tomsk Polytechnic University*, 2006, vol. 309, no. 7, pp. 142–145.

5. Gruber T. Collective knowledge systems: where the Social Web meets the Semantic Web. *Journal of Web Semantics*, 2008, vol. 6, no. 1, pp. 4–13.

6. Lande D.V. *Poisk znanii v Internet* [Search of knowledge on the Internet]. Moscow, Dialektika Publ., 2005. 272 p. ISBN 5-8459-0764-0.

7. Rogushina Yu.V., Grishanova A.Yu. [Means of intellectualization of the search for information resources in the Internet]. *Tezisy VII Mezhdunarodnoi konferentsii "Intellektual'nyi analiz informatsii IAI-2007"* [Proceedings of the VII International Conference "Intellectual Analysis of Information IAI-2007"]. Kiev, 2007, pp. 322–331. (In Russian).

8. Yi L., Liu B. Web page cleaning for web mining through feature weighting. *Proceedings of Eighteenth International Joint Conference on Artificial Intelligence (IJCAI-03)*, Acapulco, Mexico, 2003, pp. 43–48.

9. Gladun A., Rogushina J., Shtonda V. Ontological approach to domain knowledge representation for informational retrieval in multiagent systems. *Information Theories and Applications*, 2006, vol. 13, no. 4, pp. 354–362.

10. Gruber T. Toward principles for the design of ontologies used for knowledge sharing. *International Journal Human-Computer Studies*, 1995, vol. 43, pp. 907–928.

11. Goto K. *Web redesign*. Indiana, New riders, 2002 (Russ. ed.: Goto K. *Veb-redizain: kniga Kelli Goto i Emili Kotler*. St. Petersburg, Simvol-Plyus Publ., 2003. 376 p.).

12. Kupriyanovsky V.P., Yartsev D.I., Kharitonov A.A., Utkin N.A., Nikolaev D.E., Drozhzhinov V.I., Namiot D.E., Volokitin Y.I. Semantika, metadannye i ontologii v prilozheniyakh dlya umnogo goroda – novye standarty BSI [Semantics, metadata and ontologies in smart city applications - new BSI standards]. *International Journal of Open Information Technologies*, 2017, vol. 5, no. 6, pp. 94–108.

13. Gorshkov S.S. *Vvedenie v ontologicheskoe modelirovanie* [Introduction to ontological modeling]. LLC TriniData Publ., 2016. 165 p.

14. Khovanov N.V. *Analiz i sintez pokazatelei pri informatsionnom defitsite* [Analysis and synthesis of indicators in the information deficit]. St. Petersburg, St. Petersburg University Publ., 1996. 196 p.

15. *Hovanov N., Yudaeva M., Hovanov K.* Multicriteria estimation of probabilities on basis of expert non-numeric, non-exact and non-complete knowledge. *European Journal of Operational Research*, 2009, vol. 195, iss. 3, pp. 857–863.

Для цитирования:

Воробьев В.И., Солдаткина А.А. Метод онтологического анализа web-ресурса на основе метаданных // Научный вестник НГТУ. – 2018. – № 3 (72). – С. 43–58. – doi: 10.17212/1814-1196-2018-3-43-58.

For citation:

Vorobyev V.I., Soldatkina A.A. Metod ontologicheskogo analiza web-resursa na osnove metadannykh [Method of ontological analysis of a web-resource based on metadata]. *Nauchnyi vestnik Novosibirskogo gosudarstvennogo tekhnicheskogo universiteta – Science bulletin of the Novosibirsk state technical university*, 2018, no. 3 (72), pp. 43–58. doi: 10.17212/1814-1196-2018-3-43-58.