

Автоматическое построение онтологий для тематических поисковых систем

© Е.А.Рабчевский

Г.И.Булатова

Пермский государственный университет
614990, Пермь, ул.Букирева, 15.
evgeny@ranat.ru

Аннотация

Описывается модель базы знаний поисковой системы, специализирующейся под определенную предметную область. Раскрывается алгоритм автоматизированного построения онтологий с помощью анализа англоязычных текстов, содержащихся в веб ресурсах. Описывается технология RDF/A (одна из спецификаций Semantic Web).

1. Введение

Для удовлетворения своих информационных потребностей, каждый Интернет пользователь периодически посещает сайты профессиональных сообществ, подписывается и просматривает тематические рассылки и RSS подачи, ищет в поисковых системах неизвестные термины. Таким образом, у каждого профессионала выстроена своя, использующая различные Интернет технологии, система интеграции знаний в интересующей его предметной области.

Однако задачи пользователей требуют более систематизированного и настраиваемого механизма интеграции распределенных и разнородных знаний в целостную картину предметной области.

Необходимо заметить, что оригинальная спецификация WWW [1] разрабатывалась именно для решения задачи интеграции научных материалов.

Очевидно, что для эффективной интеграции данных некой предметной области из различных Интернет источников, соответствующее приложение должно работать с семантикой веб ресурсов. В этой связи, в таких приложениях актуально использование различных технологий Semantic Web [2].

2. Руководство для читателя

В третьей главе описываются базовые стандарты Semantic Web, которые используются в работе для представления знаний. Четвертая глава посвящена технологиям RDF/A и GRDDL, которые также используются в работе, и недостаточно освещены в литературе. В пятой главе формулируются сервисы, которые должна предоставлять тематическая поисковая система (специализирующаяся под определенную предметную область). Шестая глава раскрывает модель базы знаний, на основе которой строится тематическая поисковая система. В седьмой главе приводятся результаты оригинальных исследований, на основе которых выработан алгоритм, позволяющий автоматизировать процесс построения основного элемента базы знаний тематической поисковой системы — онтологии. Восьмая глава описывает построение указанной онтологии без использования средств автоматизации.

3. Представление знаний в Semantic Web

В Интернет используется множество языков представления данных, основанные на XML. В рамках проекта Semantic Web консорциумом W3 разработан ряд спецификаций для работы с машинным представлением знаний. В частности, для представления данных, имеющих графовую структуру, на основе XML, разработан язык RDF [3] (Resource Definition Framework – Среда Описания Ресурса). RDF предоставляет средства для записи триплетов, троек данных – субъект - предикат - объект. Объект и субъект соответствуют узлам графа, а предикат или свойство - направленной дуге графа. Все элементы триплета называются RDF ресурсом и идентифицируются с помощью URI идентификаторов.

Платформа RDF активно используется для представления различных данных, в частности RSS 3.0 [4] агрегаторы новостей собирают информацию в формате RDF.

Для машинного представления различных предметных областей в Интернет, используются онтологии. Онтология – спецификация концептуализации [5], или явное, формальное описание предметной

области. Как и в объектно-ориентированном описании области знаний, онтология состоит из классов и их экземпляров, У классов и экземпляров выделяются свойства, на свойства могут накладываться логические ограничения.

Поисковой системой SWOOGLE [6] на сегодня проиндексировано свыше 10 тысяч онтологий и словарей, доступных в Веб. Онтологии используются научными сообществами – для описания терминологии [7], в электронной коммерции – для описания товаров и услуг [8], и в других приложениях Интернет. Из-за своей популярности онтологии стали использоваться и в качестве баз знаний локальных интеллектуальных систем [9].

Для описания онтологий, доступных через Веб, созданы языки RDFS [10] (RDF Schema – RDF Схема) и OWL [11] (Ontology Web Language - Язык Сетевых Онтологий). В качестве своих базовых элементов данные языки используют RDF ресурсы. RDFS используется для записи словарей, а OWL – онтологий. Сетевые онтологии предоставляют более выразительные возможности по сравнению с RDF словарями, например логические операции над классами и логические ограничения свойств.

4. Семантическая разметка, RDF/A, GRDDL

RDF графы и онтологии на их основе могут размещаться в отдельных Веб ресурсах, например файлах или RDF хранилищах, доступ к которым осуществляется через RDF сервера [12]. Также RDF графы могут встраиваться в другие XML документы, например в XHTML. Встраивание RDF данных в XHTML используется для спецификации семантики (семантической разметки) контента за счет указания ссылки на понятие во внешней семантической модели, которому соответствует часть контента.

Семантическая разметка или аннотирование представляет собой явное описание семантики контента ресурса при помощи понятий семантической модели (онтологии или словаря). Такое явное описание семантики выполняется указанием четкого соответствия между определенной частью контента ресурса и его семантикой, описанной в семантической модели.

Для семантической разметки одной из рабочих групп [13] W3 консорциума была разработана технология RDF/A [14], которая позволяет встраивать RDF данные в XHTML. RDF/A является одним из множества микроформатов [15] или диалектов языков, расширений языка HTML, в котором определяется, каким образом использовать конструкции языка HTML, чтобы интерпретировать записанный таким образом HTML код, как RDF данные.

Существуют микроформаты для записи таких словарей, как vCard, DC, RDF Calendar, RSS, GeoInfo. Все указанные словари записываются в виде RDF графов, RDF/A является микроформатом для записи непосредственно RDF синтаксиса, и может быть использован для записи терминов любых RDF словарей, например тех же vCard, DC, RDF Calendar, RSS, GeoInfo.

Ниже следует пример использования терминов словаря набора данных DC [16] (словарь DC описывает мета свойства электронных документов) в XHTML.

```
<head profile="http://www.w3.org/ 2003/g/data-view">
  <link rel="schema.DC"
    href="http://purl.org/dc/" />
  <meta name="DC.Title" xml:lang="en" lang="en"
    content="Использование терминов словаря DC в
    XHTML коде" />
</head>
```

Данный XHTML соответствует триплету, субъектом которого является URI самого ресурса, предикатом – свойство Title, описанное в словаре DC по адресу <http://purl.org/dc/>, объектом – строка "Использование терминов словаря DC в XHTML коде". Вставка такого RDF триплета в заголовок HTML страницы позволит соответствующим приложениям «понять», что название документа - "Использование терминов словаря DC в XHTML коде". При этом, это название может отличаться от того, которое представлено пользователю с помощью тега <title>. Таким образом, в XHTML можно вставлять любые RDF графы.

Использование профиля [profile=http://www.w3.org/2003/g/data-view](http://www.w3.org/2003/g/data-view) необходимо для возможности указания значения "transformation" у тега rel, что необходимо для указания ссылки на механизм GRDDL извлечения (см. следующий абзац).

Для извлечения RDF данных из различных микроформатов W3 консорциум разработал технологию GRDDL [17] (Gleaning Resource Descriptions from Dialects of Languages - Извлечение Описания Ресурса из Диалектов Языков). Для работы GRDDL-скреперов (программ, извлекающих RDF данные из XHTML) в XHTML коде необходимо указать ссылку на механизм извлечения:

```
<link rel="transformation"
  href="http://www.w3.org/2000/06/dc-extract/dc-extract.xsl"/> 11
```


Механизм извлечения основан на технологии преобразования XML документов XSLT [18], в данном случае XHTML преобразуется в RDF.

5. Интеграция знаний в масштабе Интернет

Большинство материалов в Веб представляется в текстовом виде (в данной статье не обсуждается эффективность текстового представления знаний). Под интеграцией знаний масштаба всей сети понимаем процесс, в результате которого, при наличии доступа к Веб пространству, пользователь (или некое приложение) получают целостную картину определенной предметной области, притом, что отдельные знания потенциально содержатся в тексте отдельных ресурсов, распределенных в сети.

Это определение предполагает сервисы, которые должны предоставляться интегрирующим приложением:

- проверка знаний, содержащихся в различных ресурсах на противоречивость, актуальность;
- отслеживание появления новых ресурсов по соответствующей тематике;
- выявление в ресурсах новых знаний, и вписывание их в имеющуюся картину знаний;
- определение оригинального источника результатов, опубликованных в ресурсе;
- поиск Веб ресурсов с учетом семантики поискового запроса;
- объяснение в обучающей экспертной системе для данной предметной области [19].

Такое интегрирующее приложение назовем тематической поисковой системой. Для обеспечения данных сервисов интегрирующее приложение должно иметь базу знаний, представляющую предметную область, для которой предоставляются данные сервисы. В качестве основы для базы знаний интегрирующего приложения предлагаем использовать сетевую онтологию. Ряд работ [20] показал эффективность использования онтологий для решения подобных задач.

Для решения подобных задач разработчики используют лингвистические онтологии [21], которые помимо представления предметной области (формальные онтологии [22]), отражают особенности естественного языка, который используется в терминологии. Например, между понятиями онтологии могут вводиться парадигматические связи (антонимия и др.), а также отношения логического вывода (каузация, пресуппозиция и т.д.).

Данные виды отношений впервые в компьютерных лингвистических ресурсах были введены авторами тезауруса английского языка WordNet [23]. Существует RDF представление тезауруса WordNet [24].

6. База знаний и семантический индекс

Область интеграции ограничивается базой знаний приложения, а механизмы работы приложения с базой знаний являются предметно-независимыми. Основу базы знаний составляет онтология исследуемой предметной области.

Помимо онтологии база знаний приложения содержит семантический индекс Веб пространства, кэшированные Веб ресурсы, относящиеся к предметной области интеграции, которые помимо самого контента содержат его семантическую аннотацию согласно онтологии.

Для семантической разметки кэшированных ресурсов авторы используют технологию RDF/A. Открытость этого стандарта позволит использовать семантически размеченные ресурсы не только во внутренней работе приложения, но и предоставлять веб сервис семантической разметки ресурсов, относящихся к предметной области интеграции, другим информационным системам или авторам ресурсов.

Структурная организация базы знания представлена на следующем рисунке.



Семантически размеченный ресурс хранится в RDF/A части семантического индекса, а его семантика описывается при помощи ссылки на соответствующие части семантической модели. XHTML часть индекса подобна индексу традиционных поисковых систем.

7. Автоматическое построение онтологии

Интеллектуальные системы на основе онтологий показали на практике свою эффективность, однако построение онтологии требует экспертных знаний в исследуемой предметной области и

занимает существенный объем времени, поэтому актуальной задачей является автоматизация процесса построения онтологии. Для этого предлагается использовать текстовое содержание массива Веб ресурсов описательного характера определенной тематики.

Базовой является задача разработки алгоритма автоматического построения семантической карты веб ресурса с помощью анализа его текста. Семантической картой ресурса назовем отображение контента Веб ресурса в концептуализацию его содержания, представленную в виде OWL онтологии.

Для решения данной задачи был сформирован корпус англоязычных текстов, относящихся к теме Semantic Web. Ресурсы корпуса представляют собой спецификации технологий Semantic Web с сайта W3 консорциума.

Алгоритм исследовался для определенной предметной области, что объясняется профессиональными интересами автора, а также возможностью последующей оценки полученного метода сравнением результатов с онтологией, полученной с помощью экспертных знаний (параллельно с данными исследования автор анализировал выбранный корпус и создавал онтологию данной предметной области без средств автоматизации).

7.1 Автоматическое построение семантической карты ресурса

Семантическая карта ресурса строится на основе особенностей языка, которые позволяют вытягивать семантические конструкции из текста. Исследования проводились следующим образом:

- формировался набор пар «текст – конструкция языка OWL»;
- по набору выявленных пар «текст – OWL конструкция» выявлялись правила, позволяющие автоматизировать процесс отображения текста в соответствующую OWL конструкцию;

Семантическая карта строится в два этапа, на первом строится формальная семантическая OWL конструкция, на втором происходит привязка полученной конструкции к конкретной предметной области.

Сформулируем правила, использующие синтаксис языка. Правила синтаксического уровня, выявляют семантику на основе принципов построения словосочетаний и предложений. Правила формулируются, как конструкции из различных частей речи, частей предложения, предлогов и союзов, а также конкретных слов. Дополнительно вводится понятие предмета – сущности, о которой говорится в предложении, предмет может состоять из нескольких слов. Понятие предмета также используется для формулировки правил.

Рассмотрим несколько правил:

- «Сложный предмет» или «noun1 + noun2» (два подряд идущих существительных), например словосочетание «ontology editor».

Проанализируем данный пример. Можно предположить существует целый класс абстрактных редакторов – Editor. Этот класс характеризуется тем, что все его экземпляры обладают неким характерным для этого класса свойством. В данном случае, это то, что они все что-либо редактируют. Назовем это характерное свойство `mainPropertyOfEditor`. Доменом этого свойства является класс Editor. Определим диапазон этого свойства, как класс `RangeOfMainPropertyOfEditor`. Выделим класс `OntologyEditor`, который будет подклассом класса Editor. При этом значение свойства `mainPropertyOfEditor` для подкласса `OntologyEditor` имеет строго определенное значение – экземпляр класса `RangeOfMainPropertyOfEditor`, индивид `Ontology`. Данные утверждения можно представить следующим OWL кодом:

```
<owl:Class rdf:ID="Editor">
  <rdfs:comment
rdf:datatype="http://www.w3.org/2001/XMLSchema#string"
    >класс абстрактных редакторов</rdfs:comment>
</owl:Class>
<owl:Class
rdf:ID="RangeOfMainPropertyOfEditor">
  <rdfs:comment
rdf:datatype="http://www.w3.org/2001/XMLSchema#string"
    >диапазон характерного свойства редактора
(редактируемый объект)</rdfs:comment>
</owl:Class>
<owl:Class rdf:ID="OntologyEditor">
  <rdfs:subClassOf>
    <owl:Restriction>
      <owl:onProperty>
        <owl:ObjectProperty
rdf:ID="MainPropertyOfEditor"/>
      </owl:onProperty>
      <owl:hasValue>
        <RangeOfMainPropertyOfEditor
rdf:ID="Ontology"/>
      </owl:hasValue>
    </owl:Restriction>
  </rdfs:subClassOf>
```

```
<rdfs:comment
rdf:datatype="http://www.w3.org/2001/XMLSchema#st
ring"
  >класс редакторов онтологий</rdfs:comment>
  <rdfs:subClassOf rdf:resource="#Editor"/>
</owl:Class>
<owl:ObjectProperty
rdf:about="#MainPropertyOfEditor">
  <rdfs:domain rdf:resource="#Editor"/>
  <rdfs:range
rdf:resource="#RangeOfMainPropertyOfEditor"/>
  <rdfs:comment
rdf:datatype="http://www.w3.org/2001/XMLSchema#st
ring"
    >характерное свойство редактора
(редактирует)</rdfs:comment>
  </owl:ObjectProperty>
```

- «Предмет с определением» или «adjective + subject», например словосочетание «abstract syntax». Для записи соответствующего OWL кода необходимо провести рассуждения, аналогичные приведенным в предыдущем примере:

```
<owl:Class rdf:ID="Syntax"/>
<owl:Class rdf:ID="AbstractSyntax">
  <rdfs:subClassOf rdf:resource="#Syntax"/>
  <rdfs:subClassOf>
    <owl:Restriction>
      <owl:hasValue>
        <RangeOfMainPropertyOfAbstract
rdf:ID="Abstract">
          <rdfs:comment
rdf:datatype="http://www.w3.org/2001/XMLSchema#st
ring"
            >абстрактный - экземпляр класса
уровень детализации</rdfs:comment>
        </RangeOfMainPropertyOfAbstract>
      </owl:hasValue>
    <owl:onProperty>
      <owl:ObjectProperty
rdf:ID="mainPropertyOfAbstract"/>
        </owl:onProperty>
    </owl:Restriction>
  </rdfs:subClassOf>
</owl:Class>
```

```
<owl:Class
rdf:ID="RangeOfMainPropertyOfAbstract">
  <rdfs:comment
rdf:datatype="http://www.w3.org/2001/XMLSchema#st
ring"
    >диапазон                характерного                свойства
прилагательного                абстрактный                (уровень
детализации)</rdfs:comment>
  </owl:Class>
  <owl:ObjectProperty
rdf:about="#mainPropertyOfAbstract">
    <rdfs:comment
rdf:datatype="http://www.w3.org/2001/XMLSchema#st
ring"
      >главное свойство определения абстрактный
(уровень детализации)</rdfs:comment>
    <rdfs:range
rdf:resource="#RangeOfMainPropertyOfAbstract"/>
  </owl:ObjectProperty>
```

- Простое предложение, subject1 + verb + preposition + subject2 (подлежащее, сказуемое, предлог, дополнение), например «Ontology's incorporate information about classes». Что можно представить следующим OWL кодом:

```
<owl:Class rdf:ID="Ontology"/>
<owl:Class rdf:ID="Class"/>
<owl:ObjectProperty
rdf:ID="incorporateInformationAbout">
  <rdfs:range rdf:resource="#Class"/>
  <rdfs:domain rdf:resource="#Ontology"/>
</owl:ObjectProperty>
```

- subject1 + are + subject2 + that + verb + preposition + subject3 (подлежащее, are/is, дополнение, that, сказуемое, предлог, дополнение), например предложение «Decision Engineering is an emerging discipline that focuses on developing tools». Что можно представить следующим OWL кодом:

```
<owl:Class rdf:ID="EmergingDiscipline"/>
<owl:Class rdf:ID="DecisionEngineering">
  <rdfs:subClassOf>
    <owl:Restriction>
      <owl:onProperty>
        <owl:ObjectProperty
rdf:ID="focusesOn"/>
      </owl:onProperty>
```

```

    <owl:hasValue>
      <RangeOfFocusesOnProperty
rdf:ID="DevelopingTools"/>
    </owl:hasValue>
  </owl:Restriction>
</rdfs:subClassOf>
<rdfs:subClassOf
rdf:resource="#EmergingDiscipline"/>
</owl:Class>
<owl:Class rdf:ID="RangeOfFocusesOnProperty"/>
<owl:ObjectProperty rdf:about="#focusesOn">
  <rdfs:domain
rdf:resource="#EmergingDiscipline"/>
  <rdfs:range
rdf:resource="#RangeOfFocusesOnProperty"/>
</owl:ObjectProperty>

```

Отдельно выделяются правила, которые сами не строят семантическую конструкцию, но определяют, каким образом (к каким словам) применять правила, непосредственно выявляющие семантические конструкции. Например, правило «Если сложный предмет состоит из трех и более простых, то нужно применять правило «noun1 + noun2» начиная с конца».

Рассмотрим более подробно правило «adjective + subject», в котором введены свойство `mainPropertyOfAbstract` и класс `RangeOfMainPropertyOfAbstract`. Данные конструкции введены чисто формально, используя некие законы языка, однако данное свойство и класс имеют определенную семантику. Так определение `Abstract` характеризует некую особенность предмета `Syntax`. В данном случае эту особенность можно назвать, например как «степень детализации».

Если же подходить к анализу данного словосочетания с учетом семантики, указанные свойство и класс назывались бы «имеетСтепеньДетализации» и «СтепеньДетализации» соответственно.

Задача преобразования формальных семантических конструкций в конструкции, привязанные к семантике конкретной предметной области, на данный момент автором не решена. Автор считает, что для решения данной задачи требуется источник знаний со структурой подобной таблице, приведенной ниже:

Слово	Характерное свойство
Abstract	Степень детализации
Editor	Редактирует

Предполагается представить данный источник знаний в виде RDF представления WordNet подобного ресурса [24] компьютерной лингвистики.

Для решения подобной задачи предполагается получить правила, которые позволили бы выявить данную информацию, на основе статистики совместного использования слов содержащих название понятия и его семантику.

Для того чтобы привязать полученную семантическую модель к интересующей предметной области, используется словарь соответствующей тематики. В итоговой онтологии фиксируются только те семантические конструкции, в которых участвуют термины из словаря предметной области. Словарь может создаваться экспертом или автоматически на основе статистических методов классификации.

7.2 Анализ массива ресурсов

При программном построении онтологии по контенту массива Веб ресурсов используется словарь предметной области. Это позволяет отфильтровать из общего набора полученных семантических конструкций лишь те, которые играют существенную роль в онтологической картине предметной области.

Для фильтрации не устоявшихся терминов или отношений, употребленных в неверном контексте, используются статистические методы, то есть при принятии решения о включении того или иного термина или отношения в онтологию, система должна анализировать степень употребления новой концепции в:

- Веб пространстве в целом;
- на узлах заданной степени доверия;
- вместе с терминами заданной степени важности.

Степень доверия узла определяется количеством ссылок в базе знаний на данный узел. Степень важности термина определяется количеством отношений, в которых участвует термин в базе знаний. Предварительный статистический анализ позволит управлять границами исследуемой предметной области.

Семантическая карта ресурса представляет собой онтологию, представленную в виде RDF графа. Однако после построения граф еще ни как не связан с предметной областью и нет гарантий, что информация ресурса, а значит и его семантическая карта с понятиями и отношениями, соответствуют реальной картине исследуемой предметной области.

Поэтому, автоматически полученная онтология требует экспертной оценки (модерация). Интерфейс модерации должен предоставлять эксперту графическое средство, обеспечивающее быстрый переход к содержанию Веб ресурсов от их семантической модели и наоборот. Также необходим механизм трансляции семантической модели модулируемого ресурса в онтологию базы знаний и редактирования онтологии.

8. Онтология Semantic Web

Для оценки работы программного компонента автоматического построения онтологии авторы создают онтологию предметной области «Semantic Web» без средств автоматизации. Онтология создается на основе того же корпуса англоязычных тестов, из которого выявляются правила построения семантической карты ресурса. Тексты корпуса исследовались следующим образом:

- выявлялись понятия предметной области, и обозначающие их термины на русском и английском языках;
- каждое понятие дополнялось экспертным определением;
- выявлялись триплеты, содержащие найденные понятия;
- для каждого понятия и триплета фиксировался ресурс-источник.

На данный момент, выявлено 70 понятий и 260 их содержащих триплетов. В дальнейшем, после реализации компонента построения семантической карты ресурса, планируется оценить его работу, обработав исследуемый корпус с помощью компонента и сравнив результаты с полученным в ручном режиме массивом триплетов. Создан RDF словарь для хранения полученных результатов.

Авторы благодарят профессора, заведующего кафедры КСиТ Пермского Государственного Университета, М.А. Марценюка за обсуждение работы.

Литература

- [1] Tim Berners-Lee, World Wide Web: Proposal for HyperText Project. 1990. // <http://www.w3.org/Proposal.html>
- [2] Сообщество Semantic Web // <http://www.w3.org/2001/sw>
- [3] Платформа RDF <http://www.w3.org/RDF/>
- [4] Aaron Swartz, Спецификация RSS 3.0 // <http://www.aaronsw.com/weblog/000574>
- [5] Gruber, T.R. (1993) A translation approach to portable ontology specifications. Knowledge Acquisition. Vol. 5.

- [6] Swoogle - Semantic Web Search Engine. // <http://swoogle.umbc.edu/>
- [7] Е.М. Бениаминов "Алгебраические методы в теории баз данных и представлении знаний". М.: Научный мир, 2003.
- [8] Реестр товаров и услуг ООН. // <http://www.unspsc.org/>.
- [9] Рабчевский Е.А., Архипов Е.С., Проектирование экспертных систем технической поддержки на основе онтологий // Интеллектуальные системы и компьютерные науки. Материалы IX международной конференции (23-27 октября 2006 г., г. Москва). – Москва: МГУ 2006.
- [10] RDF Schema 1.0, Язык описания RDF словарей. Рекомендация W3C 10 Февраля 2004. // <http://www.w3.org/TR/rdf-schema/>
- [11] Язык OWL. // <http://www.w3.org/2004/OWL/>
- [12] Joseki - A SPARQL Server for Jena. // <http://www.joseki.org/>
- [13] Рабочая группа лучшей практики и развертывания Семантического Веба. // <http://www.w3.org/2001/sw/BestPractices/>
- [14] Встраивание RDF в XHTML RDFa. Рабочий документ W3C 12 марта 2007. // <http://www.w3.org/TR/xhtml-rdfa-primer>
- [15] Сообщество пользователей микроформатов. // <http://microformats.org/>
- [16] Метаданные Dublin Core для нахождения ресурсов, RFC2413
- [17] Рабочая группа GRDDL <http://www.w3.org/2001/sw/grddl-wg/>
- [18] Семейство Расширяемых Языков Таблиц Стилей (XSL). // <http://www.w3.org/Style/XSL/>
- [19] Гадиагулин Р., Оболочка экспертных систем XG# 1.0: Подход к автоматизированному извлечению онтологий и их применению в компоненте объяснения // Интеллектуальные системы и компьютерные науки. Материалы IX международной конференции (23-27 октября 2006 г., г. Москва). – Москва: МГУ 2006.
- [20] А.Н. Бездушный, Э.А. Гаврилова, В.А. Серебряков, А.В. Шкотин МЕСТО ОНТОЛОГИЙ В ЕДИНОЙ ИНТЕГРИРОВАННОЙ СИСТЕМЕ РАН // СОВРЕМЕННЫЕ ТЕХНОЛОГИИ В ИНФОРМАЦИОННОМ ОБЕСПЕЧЕНИИ НАУКИ. Сборник научных трудов (Москва 2003)
- [21] Добров Б.В., Лукашевич Н.В., Синицын М.Н., Шапкин В.Н., Разработка лингвистической онтологии для автоматического индексирования текстов по естественным наукам // Электронные библиотеки: перспективные методы и технологии, электронные коллекции. Труды седьмой Всероссийской научной конференции (3-7 октября 2005г., г.Ярославль). - Ярославль: ЯрГУ, 2005.

- [22] Guarino N., Some Ontological Principles for Designing Upper Level Lexical Resources. // Proceedings of First International Conference on Language Resources and Evaluation, 1998.
- [23] C. Fellbaum. WordNet: An Electronic Lexical Database. MIT Press, 1998.
- [24] RDF/OWL представление WordNet, Рабочий документ W3C 19 Июня 2006 <http://www.w3.org/TR/wordnet-rdf/>

Automatic ontologies construction for custom search engines

Evgeny Rabchevsky, Gulnara Bulatova

This paper describe knowledge base model for search engine that are customized for some domain. It defines algorithm for automatic ontology construction by English-speaking web resources text analysis. There are quoted RDF/A technology (some of Semantic Web specification).