Семантический Веб
Лекция № 8 курса
«Алгоритмы для Интернета»

Юрий Лифшиц*

16 ноября 2006 г.

Содержание

1. История и мотивация

Семантический Веб — новая концепция развития Веба и сети Интернет, принятая и продвигаемая W3C (World Wide Web Consortium). Эта организация разрабатывает и внедряет технологические стандарты для Всемирной паутины.

Когда сейчас слово «Интернет» употребляется в обиходе, то чаще всего имеется в виду Всемирная паутина и доступная через нее информация, а не сама физическая сеть компьютеров.

Интернет — всемирная система объединенных компьютерных сетей, построенная на использовании протоколов (TCP/IP) для связи и маршрутизации пакетов данных.

Веб — глобальное информационное пространство, основанное на физической инфраструктуре Интернета и протоколе передачи данных HTTP. При его создании в 1989 Тимом Бернерсом-Ли предполагалась не только человеческая взаимосвязь, но и участие компьютеров в осмысленном манипулировании информацией. Основной помехой для этого является тот факт, что большинство данных в Вебе хранится в форме, рассчитанной на восприятие человеком. Таким образом, даже при учете того, что информация получена из какой-либо базы, ее структура не очевидна роботу. Компьютер способен умело парсить веб-страницу, но, вообще говоря, у него нет надежного способа для извлечения семантики документа. Семантический Веб как раз имеет своей целью наверстать это упущение.

*Законспектировал Андрей Клебанов.

Семантический Веб — это не какая-то отдельная сеть, а расширение уже существующей, такое, что в ней информация снабжена точно определенным значением, что позволит человеку и машине успешней взаимодействовать. Первые этапы на пути к «вплетению» Семантической Сети в структуру имеющейся Сети уже осуществляются полным ходом. В ближайшем будущем данные разработки возвестят о новых значительных функциональных возможностях, когда машины станут намного лучше обрабатывать и «понимать» те данные, которые сейчас они просто показывают на экране.

1.1. Сценарии будущего

Агент — программа, работающая без непосредственного управления со стороны человека или другого постоянного контроля, созданная для достижения целей, поставленных перед ней пользователем. Обычно агенты собирают, фильтруют и обрабатывают информацию, найденную в Сети, иногда с дополнительной помощью со стороны других агентов.

Семантический Веб привнесет структуру в смысловое содержание веб-страниц, тем самым создав среду, в которой агенты, переходя со страницы на страницу, смогут без особого труда выполнять замысловатые запросы пользователя.

При условии существования подобных программ возможны следующие запросы человека своему агенту:

- Закажи для меня эту книгу в ближайшей библиотеке.
- Посмотри на расписание электричек и мое расписание и выбери билеты в театр, чтобы я мог успеть после работы.
- Скажи мне, какое вино нужно купить к каждому из блюд в этом меню. И кстати, я не люблю Сотерн.
- Микроволновка, сходи на сайт производителя и загрузи оптимальные параметры подогрева.

Можно выделить несколько вариантов использования Семантического Веба.

Семантический поиск. Поисковая система сможет выдавать только те сайты, где упоминается в точности искомое понятие, а не произвольные страницы, в тексте которых встретилось данное многозначное ключевое слово. Сегодняшние поисковые системы зачастую выдают бесчисленное множество совершенно не относящихся к запросу страниц, обрекая пользователя на длительный ручной отбор материала. Например, если вы ввели для поиска слово «cook», то компьютеру совершенно непонятно, имеете ли вы в виду повара, хотите ли найти информацию о рецептах приготовления пищи, или же вам нужно какое-то место, человек или компания или еще что-либо, в чьем имени или названии встречается слово «cook». Вся проблема в том, что для компьютера слово «cook» не имеет четкого смысла, или другими словами, семантического содержания.

Объединение знаний (интеграция баз данных). Как было сказано выше, отсутствие семантики ведет к тому, что компьютеру не ясно, как поступить с информацией из базы данных даже при условии того, что он знает названия всех столбцов полученной таблицы. Семантический Веб, именуя всякое понятие просто с помощью URI-идентификатора, даст возможность каждому выражать новые понятия, которые он изобретает, с минимальными усилиями. Его универсальный логический язык позволит постепенно связать все эти понятия в универсальную Сеть. Эта структура сделает знания и достижения человечества доступными для анализа программными агентами и предложит нам новый класс средств, с помощью которых мы сможем вместе жить, работать и учиться. В частности, работа *вопросо-ответных систем* станет значительно эффективней.

Всепроникающие вычисления (ubiquitous computing). На следующем этапе своего развития Семантическая Сеть уже вырвется из виртуальной области и расширит сферу своего влияния на физический мир, поскольку URI-идентификаторы могут указывать на что угодно, в частности, и на физические объекты, такие как сотовый телефон или телевизор. Эти устройства смогут рекламировать свои функциональные возможности (что они умеют делать и каким образом ими можно управлять) практически так же, как это делают программные агенты. Например, так называемая домашняя автоматизация требует сейчас тщательной настройки всех устройств для их совместной работы. Семантическое же описание

возможностей и порядка функционирования этих устройств позволит достичь той же автоматизации, но уже с минимальным вмешательством человека.

1.2. Хронология

- 1994: Создание W3C. Консорциум разработал стандарты: HTML, URL, XML, HTTP, PNG, SVG, CSS.
- 1998: Тим Бернерс-Ли публикует план Семантического Веба (Semantic Web Road map).
- 1999: W3C создает группы по проектированию Семантического Веба, публикуется первая версия RDF.
- 2000: Американские военные начинают исследования по описанию онтологий (DAML+OIL project).
- 2001: В журнале Scientific American публикуется описание Семантического Веба.
- 2004: Выпущена новая версия RDF, представлен язык описания онтологий OWL.
- 2006: Представлена версия языка запросов SPARQL (candidate recommendation).

В разработке проекта участвуют: HP, Mozilla, IBM, MIT, Stanford, ...

2. Архитектура Семантического Веба

Что нам нужно сделать, чтобы можно было осуществить вышеприведенные сценарии будущего? Для простоты будем отталкиваться от того, что было разработано при создании обычного Веба: протокол передачи данных, язык разметки страниц и браузер.

Таким образом, наивный план будет представлять собой следующее:

1. Разработать язык огромной выразительной силы, на котором можно описать все знания человека и который был бы понятен компьютерам.
2. Перевести все сайты на этот язык.
3. Написать программы, работающие со знаниями на этом языке (обработка запросов, логический вывод).

К сожалению, первые два пункта представляются невыполнимыми. Поэтому рассмотрим более тонкое решение, предложенное Тимом Бернерсом-Ли.

2.1. Общая архитектура

В отличие от подхода, использующего искусственный интеллект для обучения компьютеров поведению людей, его идея заключается в поэтапной и распределенной разработке языка, способного выражать информацию в понятной машине форме. Таким образом, цель Семантического Веба — создание языка, на котором можно будет описать как данные, так и правила рассуждений об этих данных, так что правила вывода, существующие в какой-либо системе представления знаний, можно будет экспортировать в Веб.

Чтобы определить язык, необходимо задать его синтаксис и семантику.

Синтаксис — набор правил построения фраз языка, позволяющий определить корректные предложения в этом языке. Основным инструментом синтаксиса является наличие правил проверки, позволяющих судить о том, удовлетворяет ли текст синтаксису или нет.

Семантика — система правил истолкования отдельных языковых конструкций. Семантика определяет смысловое значение предложений языка.

Примером языка с синтаксисом, но без семантики, является XML, а примером семантики без синтаксиса — человеческая речь, поэтому программам так трудно разобраться, где что.

Тим Бернерс-Ли предложил *отдельно* разрабатывать синтаксис и семантику языка описания всех знаний человечества:

- *RDF* (Resource Description Framework) — язык, отвечающий за синтаксис документов Семантического Веба. В нем широко используются ссылки на онтологии для определения смысла слов.
- *OWL* (Ontology Web Language) — язык описания онтологий.

Онтология — описание классов объектов, их свойств и взаимоотношений для какой-то предметной области (домена).

RDF и OWL будут детальнее рассмотрены в дальнейшем.

Таким образом, план Тима Бернерса-Ли требует последовательно разработать:

1. Синтаксис для представления знаний, использующий ссылки на онтологии (сделано: RDF).
2. Язык описания онтологий (сделано: OWL).
3. Язык описания веб-сервисов (начато: WSDL, OWL-S).

   В настоящее время уже существует множество автоматизированных веб-сервисов безо всякой семантики, однако у других программ, таких как агенты, нет никакого способа разыскать в сети подобную программу, выполняющую ту или иную функцию. Этот процесс, называемый обнаружением сервисов, станет возможным лишь после появления единого языка, позволяющего описывать сервисы, чтобы агенты могли «понимать», что позволяет делать данный сервис и каким образом им пользоваться. Сервисы и агенты могут рекламировать выполняемые ими функции, например, занося подобные описания в справочники, подобные «Желтым Страницам».
4. Инструменты чтения и разработки документов Семантического Веба (начато: Jena, Haystack, Protege).

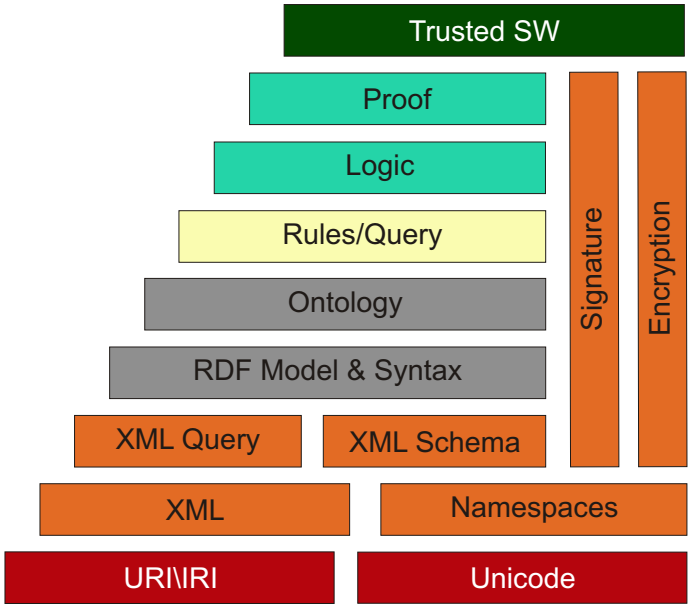
   Главный минус концепции Семантического Веба — сложность внедрения. Формат RDF был разработан людьми с академическим образованием и изначально не был рассчитан на применение рядовыми пользователями Интернета. Даже многим веб-мастерам и программистам бывает сложно освоить RDF и OWL. Но, несмотря на это, Тим Бернерс-Ли утверждает, что в будущем никаких специальных знаний для создания страниц не потребуется.
5. Язык запросов к знаниям, записанным в RDF (начато: SPARQL).

   SPARQL — новый язык запросов для быстрого доступа к данным RDF. Используя обычный протокол и язык SPARQL, приложения могут анализировать RDF-описания ресурсов и получать из сети нужную информацию.
6. Логический вывод знаний (не сделано).
7. Семантическая поисковая система (начато: SHOE).
8. Агенты Семантического Веба (не сделано).

Ключевым аспектом технологий Семантического Веба является их многоуровневая архитектура, которая показана на рисунке.

Верхние уровни зависят от нижних, которые в свою очередь могут использоваться независимо.

В самом низу находится Unicode — стандарт кодирования символов, позволяющий представить знаки практически всех письменных языков. Рядом с ним находится URI, гарантирующий то, что каждое понятие, используемое в документе — это не просто слово, а нечто, привязанное к единому определению, которое каждый желающий может найти в Сети. Иметь подобный механизм в самом низу необходимо для того, чтобы предоставить каждому человеку универсальный способ описания ресурсов.



Пирог уровней Семантического Веба, представленный Тимом Бернерсом-Ли на конференции XML 2000

XML используется в качестве синтаксиса документов и совместно с пространством имен может использоваться для выражения информации и обмена ею между программами. XML Schema используется для выражения набора правил, которым должен удовлетворять XML-документ, чтобы быть признанным корректным. Протокол HTTP и язык HTML лежат где-то в области этого уровня, но воспринимаемые человеком документы и их передача в значительной степени отличаются от материала, ориентированного преимущественно на машины, поэтому их обычно не показывают на этой диаграмме.

Пропуская RDF и OWL, перейдем к верхним уровням диаграммы. Логический вывод (Logic) используется для обеспечения связности и корректности информации, а также для получения новых данных. Доказательства (Proof) отслеживают и объясняют шаги логического вывода. Заслуживающий доверия Семантический Веб (Trusted SW) — средства, выполняющие аутентификацию, проверку достоверности информации, надежности сервисов и агентов.

2.2. RDF: синтаксис Семантического Веба

Язык XML дает возможность пользователям создавать документы произвольной структуры, однако данный язык ничего не говорит о том, что означает эта структура. Смысл выражается посредством языка RDF, который кодирует его с помощью деревьев глубины три (Notation3), где каждое дерево состоит из субъекта (подлежащее), свойства (сказуемое) и объекта (дополнение). Объект можно назвать функцией свойства от субъекта. Например, утверждение «Небо голубого цвета» в RDF-терминологии можно представить следующим образом: субъект — «небо», свойство — «иметь цвет», объект — «голубой». Для идентификации субъектов, свойств и объектов в RDF используются URI. В начале любого RDF документа идет список ссылок на онтологии. Таким образом, каждая вершина может быть задана строкой или ссылкой на объект из какой-либо онтологии. Вершины могут иметь дополнительные уточняющие квалификаторы.

Частным случаем формата RDF является формат RSS. RSS — семейство XML-форматов, предназна-

ченных для описания лент новостей, анонсов статей, изменений в блогах и т. п. Информация из различных источников, представленная в формате RSS, может быть собрана, обработана и представлена пользователю в удобном для него виде специальными программами-агрегаторами. Под RSS может пониматься: Rich Site Summary — богатая сводка сайта; RDF Site Summary — сводка сайта с применением инфраструктуры описания ресурсов; Really Simple Syndication — очень простая синдикация.

Рассмотрим пример документа в формате RDF.

```
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
     xmlns:dc="http://purl.org/dc/elements/1.1/"
     xmlns:geo="http://www.w3.org/2003/01/geo/wgs84_pos#"
     xmlns:edu="http://www.example.org/">
     <rdf:Description rdf:about="http://www.princeton.edu">
     <geo:lat>40.35</geo:lat>
     <geo:long>-74.66</geo:long>
     <edu:hasDept rdf:resource="http://www.cs.princeton.edu"
         dc:title="Department of Computer Science"/>
     </rdf:Description>
</rdf:RDF>
```

Верхний тэг означает, что все его содержимое — RDF-описание, то есть семантическая часть страницы, понятная компьютеру. Атрибут `xmlns` является ссылкой на пространство имен. После двоеточия следует название, а в кавычках URI, `rdf` — базовое пространство имен, а три остальные — дополнительные. Первое из них, `dc`, будет рассмотрено в дальнейшем, второе, `geo`, отвечает географической онтологии, а третье — образовательной. Тэг `rdf:Description` определяет тройное предложение, его атрибут `rdf:about` — субъект. Далее следуют три свойства: «широта», «долгота» и «иметь факультет». Объекты в первых двух случаях заданы не ссылкой, а строкой. В третьем случае у тэга `edu:hasDept` есть атрибут `rdf:resource`, определяющий объект. Далее, используя пространство имен Dublin Core, задается название объекта.

2.3. OWL: язык описания онтологий

Основные компоненты OWL включают классы, свойства и индивидуальные элементы.

Класс — это концепция в онтологии. Классы являются основными блоками OWL и обычно образуют таксономическую иерархию (т.е. систему подкласс-надкласс). OWL поддерживает шесть основных способов описания классов. Самый простой — класс с именем (named). Другие типы: классы пересечений (intersection), объединений (union), дополнений (complement), ограничений (restrictions) и классы перечислений (enumerated).

Индивидуальные элементы — это элементы классов. В RDF они будут объектами и субъектами.

Мир классов и индивидов был бы совершенно неинтересным, если бы мы могли только определять таксономии. Свойства позволяют нам утверждать общие факты о членах классов и особые факты об индивидах. Они включают две основные категории: *свойства-объекты*, которые связывают индивидуальные элементы между собой и *свойства-значения* (datatype properties), которые связывают индивидуальные элементы со значениями типов данных. Для определения типов данных OWL использует схему XML.

Характеристики свойств:

- Симметричность: $\forall x$, y и свойства R: $R(x, y) \Rightarrow R(y, x)$.
- Транзитивность: $\forall x$, y, z и свойства R: $R(x, y) \wedge R(y, z) \Rightarrow R(x, z)$.
- Функциональность: $\forall x$, y, z и свойства R: $R(x, y) \wedge R(x, z) \Rightarrow y = z$. Свойством функциональности обладает дата рождения человека, у каждого человека она единственна.

К классам и свойствам могут применяться различные ограничения. Например, ограничения мощности множества указывают на число связей, в которых может участвовать класс или индивидуальный элемент. Также в OWL существуют команды для склеивания (эквивалентности) классов.

3. Проекты Семантического Веба

Дублинское ядро (Dublin Core)

Дублинское ядро (Dublin Core) появилось раньше RDF, но теперь это просто аннотации (метаданные) к любым объектам, записанным на RDF с помощью онтологии DC. Цель DC — установить единый формат метаданных для облегчения поиска по автору, названию, году выпуска и т. д.

Множество метаданных Дублинского ядра (Dublin Core Metadata Element Set (DCMES)) состоит из 15 элементов: Title, Creator, Subject, Description, Publisher, Contributor, Date, Type, Format, Identifier, Source, Language, Relation, Coverage, Rights.

Рассмотрим пример.

```
<?xml version="1.0"?> <metadata
        xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
        xmlns:dc="http://purl.org/dc/elements/1.1/"
        xmlns:dcterms="http://purl.org/dc/terms/">
  <dc:title> Automated
      Theorem Proving
  </dc:title>
  <dc:creator>
      Mantsivoda Andrei
  </dc:creator>
  <dc:subject xsi:type=dcterms:UDC>
      681.3
  </dc:subject>
  <dc:date>
      2004-04-04
  </dc:date>
  <dc:type>
     Article
  </dc:type>
  <dc:identifier>
      http://andrei.baikal.ru/atp2004
  </dc:identifier>
</metadata>
```

Атрибут `xmlns` является ссылкой на пространство имен. Первое относится к XML, второе и есть Dublin Core, а последнее — дополнительный словарик. Заметим, что у тэга `dc:subject` есть атрибут «UDC» из дополнительного словарика.

Друг друга

Друг друга (Friend of a Friend, FOAF) — онтология характеристик личностей и человеческих взаимоотношений. Цель — снабдить домашние страницы и профили машинно-понимаемыми описаниями, объединив все социальные сети в одну глобальную базу.

Haystack

В проекте Haystack разрабатывается индивидуальная система управления информацией. Два подпроекта имеют отношения к Семантическому Вебу.

Piggy-Bank (Re:Search) сохраняет историю навигации в RDF формате, позволяя проводить «умный» поиск по материалам уже просмотренных страниц.

Haystack может использоваться как семантический браузер по документам, снабженным RDF-описаниями, а также производить обработку таких страниц.

Другие проекты

- Поисковая система SHOE (http://www.cs.umd.edu/projects/plus/SHOE/search/): поиск в Семантическом Вебе.
- Jena (http://jena.sourceforge.net): среда разработки приложений для Семантического Веба, включает исполнитель SPARQL-запросов.
- Simile (http://simile.mit.edu): Семантический Веб для электронных библиотек.
- Protege (http://protege.stanford.edu): редактор онтологий из Стэнфорда.

Итоги

Семантический Веб — снабжение Интернет страниц описаниями, которые понятны компьютерам.
Описания пишутся на языке RDF со ссылками на онтологии, построенные с помощью языка OWL.
Отдельные подпроекты Семантического Веба имеют самостоятельное значение: FOAF, DC, RSS.

В полную силу Семантический Веб будет реализован тогда, когда люди создадут множество программ, которые, знакомясь с содержимым Сети из различных источников, обрабатывают полученную информацию и обмениваются результатами с другими программами. Эффективность таких программных агентов будет расти экспоненциально по мере увеличения количества доступного машинно-воспринимаемого веб-контента и автоматизированных сервисов (включая других агентов).

Источники

[1] Тим Бернерс-Ли, Джеймс Хендлер и Ора Лассила. Семантический Веб
http://ezolin.pisem.net/logic/semantic_web_rus.html

[2] Joshua Tauberer. What Is RDF?
http://www.xml.com/pub/a/2001/01/24/rdf.html

[3] Рекомендация W3C. Перевод Дмитрия Щербины. OWL, язык веб-онтологий. Руководство
http://sherdim.rsu.ru/pts/semantic_web/REC-owl-guide-20040210_ru.html

[4] А.В. Манцивода. Система метаописаний Dublin Core
http://teacode.com/concept/eor/dc.html

[5] Страница курса
http://logic.pdmi.ras.ru/~yura/internet.html