УДК 002.53:[510.63:681.3.016]

ИНТЕЛЛЕКТУАЛИЗАЦИЯ ОБРАБОТКИ ИНФОРМАЦИИ НА ОСНОВЕ ТЕХНОЛОГИЙ SEMANTIC WEB

С.С. Щербак
(представил д.ф.-м.н., проф. С.В. Смеляков)

В статье рассмотрены и предложены пути решения проблемы анализа и автоматической обработки информации на основе онтологического подхода с использованием технологий Semantic Web.

Введение. С каждым днем количество пользователей сети Интернет увеличивается, существуют миллионы серверов, предоставляющих различного рода электронные документы. Эффективность компьютерного анализа электронных документов оставляет желать лучшего. Становится все более очевидным отсутствие эффективных методов извлечения и формализации знаний с электронных документов для дальнейшего, с учетом смысла, анализа. Такая ситуация наблюдается как при анализе текстовых документов – неструктурированных или слабоструктурированных, так и в табличных – структурированных с помощью таблиц электронных документах. Не стоит также опускать немаловажный фактор большой гетерогенности источников электронных документов. Над решением этих проблем работают множество исследовательских коллективов. Текущее состояние интеллектуальной обработки электронных документов заключается в создании семантически интероперабельной (способной к взаимодействию) среды для “интеллектуальных” программ. Эта среда получила название – Semantic Web.

1. Web, основанный на понимании информации. На сегодняшний день информация для людей и компьютеров готовится отдельно: для людей – в виде текста, картинок и звуков, для машин – в виде специальных кодов. Semantic Web предусматривает объединение этих разных видов информации в единую структуру, где каждому элементу “человеческой” информации будет соответствовать машинный код в виде специального смыслового тэга (метаданные). Все тэги должны составлять единую иерархическую структуру RDF, на основе которой и будет работать Semantic Web. Метаданные будут в обязательном порядке включать сведения о том, как, где и кем была создана данная информация и как она структурирована. Таким образом, унифицированное представление информации в Semantic Web плюс набор механизмов “понимающих”

смысловые теги, заложенных в эту информацию обеспечат компьютерную обработку информации с учетом ее семантики.

Росту популярности и широкому внедрению технологий Semantic Web способствует стандартизация консорциумом W3C синтаксической и семантической разметки электронных документов, особенно технологий XML, RDF/RDFS и OWL, поддерживающих синтаксическую и семантическую совместимость. Кроме того, среда Semantic Web обеспечит классификацию информации, что сделает совместную работу людей и компьютеров на порядок более эффективной.

В основе Semantic Web лежат следующие концепции: расширяемый язык разметки XML; RDF – формат описания ресурсов; онтологии, определяющие термины и отношения между ними.

2. Расширяемый язык разметки XML. За последние несколько лет широкое распространение получила технология XML (eXtended Markup Language), к достоинствам которой можно отнести:

1) расширяемый язык разметки электронных документов, обеспечивающий возможность создания унифицированного представления электронных документов, их структуры на основе словаря разметочных тегов и правил составления тегов в синтаксические конструкции;

2) развитые средства синтаксического анализа унифицированного представлений электронных документов;

3) кроссплатформенность и совместимость с гипертекстовой средой Интернет;

4) возможность разметки документов произвольной структуры.

Правильно составленные XML-документы содержат сбалансированное дерево вложенных открывающих и закрывающих тегов, каждый из которых может включать в себя несколько пар “атрибут-значение”. Поскольку фиксированного словаря тегов, равно как и набора их допустимых комбинаций не существует, теги могут определяться независимо от приложения. В XML это делается с помощью определяемых пользователем словарей тегов в виде схем XML или DTD(определения типа документа), накладывающих ограничения на используемые теги и указывающих, каким образом должна быть организована их вложенность внутри документа. Схемы XML или DTD задают грамматику, которая указывает допустимые комбинации и вложения имен тегов, имен атрибутов и т.д.

Как технология Semantic Web XML обеспечивает общую синтаксическую спецификацию для представления информации. Кроме того, наличие средств для синтаксического разбора и обработки информации, выраженной в XML синтаксисе, интегрированность в средства коммуникации Интернет, позволяет обеспечить естественную среду для развития и практиче-

ского применения, рассматриваемых ниже, онтологий, выраженных в XML-синтаксисе, для обработки и обмена онтологиями в среде WWW.

Широкая поддержка и внедрение XML в Web обеспечили для приложений синтаксически интероперабельную (способную к взаимодействию) среду, позволив эффективно решать проблемы обмена информацией и межпрограммного взаимодействия. Однако унифицированное представление документов на XML ничего не говорит о том, что означает это унифицированное представление, т.е. не несет никакой семантической нагрузки.

3. Формализация и обработка знаний на основе онтологического подхода: в рамках Semantic Web Интернет рассматривается как распределенная база знаний. Для работы с распределенными знаниями в Интернет, нужны специальные методы представления и обработки, распределенных по всемирной паутине WWW знаний. Задача заключается, прежде всего, в том, чтобы адаптировать методы и средства, разработанные в искусственном интеллекте для знание-ориентированных систем, в новую проблемную область. В рамках такого подхода сегодня внимание различных исследователей привлекают онтологии, как средство построения распределенных и неоднородных систем баз знаний на основе Интернет. Вопросам, связанным с формализацией знаний, компьютерному анализу знаний посвящено множество работ [1 – 3]. Достоинствами онтологий являются их потенциальные свойства для решения таких задач, как формализация, интеграция, обмен знаниями и их повторное использование. Это заключение основывается на предположении о том, что если общая схема представления и использования знаний, – то есть онтология, – явно определена для работающих с ней “интеллектуальных” приложений как общий ресурс, то этот ресурс, возможно, разделять между “интеллектуальными” приложениями и многократно его использовать [1].

Онтология представляет собой формальное, явное описание понятий предметной области и отношений между ними, а также правила для составления новых понятий и отношений. Очень важным в данном определении является то, что онтология, кроме уже определенных понятии и отношений, содержит также правила для получения новых понятий и отношений. Учитывая, что онтология предназначена для “машинного” чтения, типы понятий и ограничений, определенных в онтологии явно определены.

Формально записанные знания в онтологии составляют семантическую основу – базу знаний, для компьютерного анализа информации, кроме того, онтологии предоставляют возможность семантического взаимодействия между “интеллектуальными” приложениями независимо от их индивидуальных особенностей, структуры информации и областей применения.

В рамках Semantic Web получили широкое распространение языки описания знаний в онтологиях, основанные на XML – RDF, RDFS, OWL.

Формат описания ресурсов представляет собой возможность выражения метаданных о ресурсах в терминах “объект-атрибут-значение”. Последовательно выраженные RDF-графы цепочек описаний метаданных позволяют выразить в “Машино-понимаемом” формате семантические описания ресурсов. Словарь терминов [понятий], используемых в семантических описаниях RDF задается с помощью схемы RDF – RDFS. Усилиями сообщества Web разработчиков и консорциума W3C была разработана более продвинутая версия RDFS – язык Web онтологий(OWL), в которую добавлена возможность более выразительного описания классов и отношений между ними.

Основанный на языке LBASE, в качестве ядра которого используется хорошо понимаемая логика первого порядка, OWL представляет собой одно из наиболее развитых средств семантического описания ресурсов. Существует три разновидности OWL – OWL Lite, OWL DL, OWL Full.

OWL Lite предназначен для тех пользователей, которых в основном интересует классификационная иерархия и простые ограничения.

OWL DL предназначен для тех пользователей, которые хотят максимум выразительности, сохраняя при этом полноту вычислений, т.е. все выражения будут гарантированно вычисляемыми и все вычисления будут завершены за конечное время.

OWL Full предназначен для пользователей, которые нуждаются в максимальной выразительности и синтаксической свободе RDF без вычислительных гарантий.

Для организации программной обработки семантических описаний используется декларативный язык запросов RDF Query, наиболее полной реализацией которого является программный обработчик Jena от Hewlett-Packard. Jena содержит реализации наиболее общих методов для работы с моделями онтологий, таких как навигация и обработка моделей онтологий в виде RDF-троек, наборов ресурсов со свойствами и т.д. Следует отметить также то, что Jena обеспечивает возможности построения модели данных онтологии, нахождения различий между моделями, интеграции онтологий и т.п.

Таким образом, на основе рассмотренных технологий Semantic Web, может быть организована автоматизированная, а для некоторых областей и автоматическая обработка информации с учетом ее смысла.

4. Модель интеллектуальной системы обработки информации, построенной на основе технологий Semantic Web. Четко определенный семантический базис предметной области позволяет организовать более “осмысленный” анализ информации в электронных документах. Во-первых, любые естественно-языковые конструкции, с помощью которых может выражаться та или иная информация, содержит в явном или неявном виде

предмет обсуждения, семантическую идентификацию которого можно осуществить благодаря наличию онтологии предметной области, кроме того, могут быть определены потенциальные взаимосвязи между объектами и идентифицированы в тексте. Во-вторых, информация в электронных документах, особенно та, которая публикуются в Интернете, часто либо структурирована, либо содержит структурированные островки информации, в виде списков, таблиц. Идентификация описания информации, в виде названий атрибутов, составляющих заголовки структурированной информации, также может быть осуществлена с помощью онтологии. Не имея онтологии, островки структурированной информации, могут быть неправильно разделены программным обработчиком на значения и описания этих значений, т.е. будут неправильно построены цепочки “атрибут-значение”, описывающие список или таблицу. Поэтому представляется целесообразным использование онтологии предметной области для организации идентификации семантических объектов и их взаимосвязей в представлении информации в электронных документах.

Идентификация семантических объектов информации определяется как процесс отображения составляющих естественно-языковых конструкций на семантические описания объектов в онтологии предметной области. Здесь одну из главных ролей, выполняет полнота описания предметной области, т.е. онтологии. Кроме того, в онтологии должны быть учтены синонимы, соответствующие тому или иному семантическому объекту. Проблема омонимии языков может быть решена путем идентификации семантических объектов и проверки на допустимость возможных взаимосвязей этих идентифицированных объектов. Таким образом, анализ электронного документа сводится к следующим, последовательно выполняемым шагам, как изображено на рис. 1. Результатом анализа является семантически размеченный документ, т.е. документ в котором выделены семантические объекты, идентифицированы основные взаимосвязи. В качестве языка семантической разметки может выступать может выступать один из языков, применяемых в Semantic Web

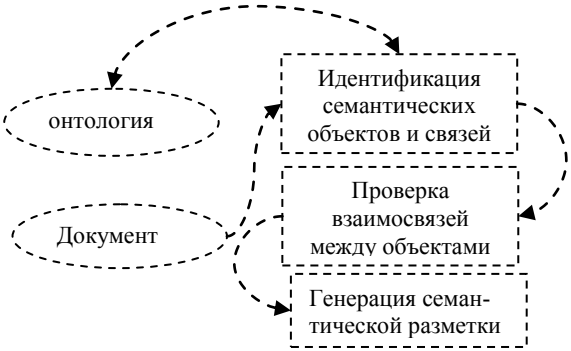

Рис. 1. Анализ документа на основе онтологии предметной области

для описания метаинформации об объектах, например, формат описания ресурсов RDF/RDFS или OWL. Наиболее целесообразным представляется генерация семантической разметки в формате, совместимом с языком описания знаний онтологии предметной области, что создаст естественную среду для интеграции полученных семантических описаний в онтологию предметной области.

Моделирование процесса анализа документов на основе технологий Semantic Web позволил построить модель программной системы обработки информации электронных документов, изображенную на рис. 2.

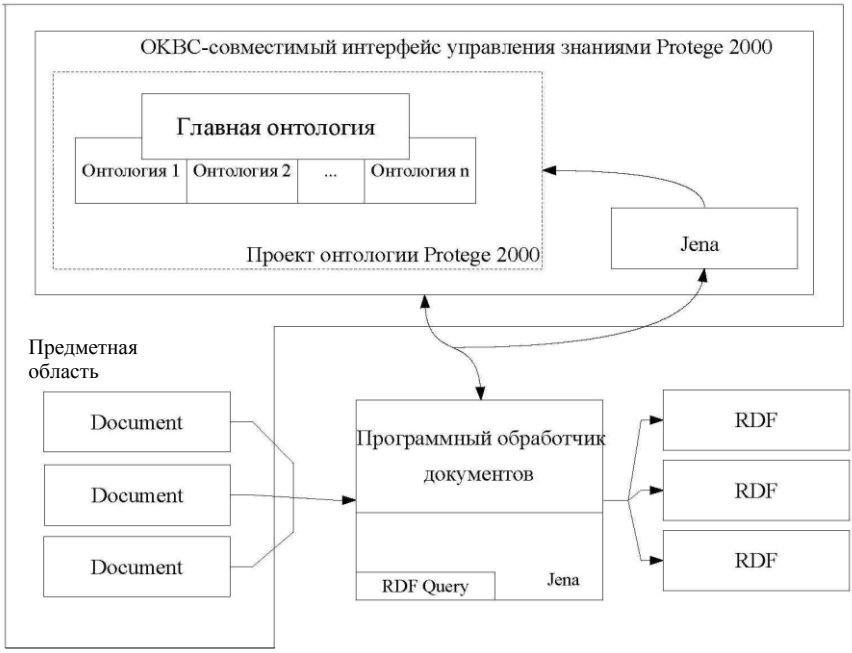

Рис. 2. Модель интеллектуальной системы обработки информации

Основу этой программной системы составляет иерархически организованный проект онтологии, состоящий из главной онтологии и онтологий более низких уровней, ответственных за решение каких-либо специфических для конкретной предметной области задач. Программный обработчик документов идентифицирует семантические объекты, выделяя синтаксические конструкции языка документа и определяя их семантические характеристик путем отображения на онтологию предметной области.

Для управления проектом онтологии представляется целесообразным применение программного комплекса Protégé-2000[4], представляющего со-

бой исключительно мощное средство для создания и поддержки онтологий, использующее OKBC – совместимый интерфейс управления знаниями[5], что позволяет Protégé-2000 использовать единый интерфейс для работы с различными языками семантической разметки. Кроме того, благодаря возможности расширения функциональности программного комплекса Protégé-2000 за счет добавления соответствующих плагинов – встраиваемых функциональных модулей, можно эффективно управлять содержимым онтологии, так использование встраиваемого плагина, построенного на основе пакета Jena от Hewlett Packard позволяет Protégé-2000 манипулировать различными онтологиями с целью интеграции, например, с проектом онтологии верхнего уровня, а также выполнять различные задачи по трансформации онтологии с одного языка описания знаний в другой, генерации различных представлений и т.д.

Выводы. В результате проведенных исследований разработан подход к анализу документов и организации автоматизированной обработки электронных документов на основе онтологий предметных областей и технологий Semantic Web, построена модель интеллектуальной системы обработки информации электронных документов, определены основные требования для успешной программной идентификации семантических объектов и их взаимосвязей.

ЛИТЕРАТУРА

1. *Enabling Technology for Knowledge Sharing / R. Neches, R. Fikes, T. Finin, T. Gruber, R. Patil, T. Senator, R. Swartout // AI Magazine. – 1999. – Vol. 12. – № 3.*
2. *Berners-Lee T. Weaving the Web. – Harper, San Francisco, 1999.*
3. *Проект Semantic Web. – [Электронный ресурс]. – Режим доступа: http://www.w3c.org/sw.*
4. *Проект Protégé. – [Электронный ресурс]. – Режим доступа: http://protégé.stanford.edu.*
5. *Open Knowledge Base Connectivity (OKBC). – [Электронный ресурс]. – Режим доступа: http://www.ai.sri.com/~okbc/.*
6. *Shelley Powers. Practical RDF. – O'Reilly, 2003. – 350 с.*
7. *RDF/XML Syntax Specification. – [Электронный ресурс]. – Режим доступа: http://www.w3.org/TR/rdf-syntax-grammar.*
8. *RDF Premier. – [Электронный ресурс]. – Режим доступа: http://www.w3.org/TR/rdf-prime/.*
9. *Спецификация языка RDFS. – [Электронный ресурс]. – Режим доступа: http://www.w3c.org/rdfs.*

Поступила 2.09.2004

ЩЕРБАК Сергей Сергеевич**, аспирант, асс. каф. искусств. интеллекта ХНУРЭ. В 2002 году окончил ХНУРЭ, магистр. Область научных интересов – онтологический подход к анализу документов. Контакт: **e-mail:** spec_sergey@ukr.net **WWW: http://ontolib.com