ОНТОЛОГИЧЕСКОЕ АННОТИРОВАНИЕ ИНФОРМАЦИОННЫХ РЕСУРСОВ ДЛЯ ОПТИМИЗАЦИИ СМЫСЛОВОГО ПОИСКА.

к.т.н., доц. Рябова Н.В., Шаламов М.А.
Харьковский национальный университет радиоэлектроники
(61166, Харьков, пр. Ленина,14, каф. ИИ, тел. (057) 70-21-337),
E-mail: mshalamov@ukr.net.

The given work is devoted search of new methods and approaches which would improve process of extraction and information integration. The given work is devoted a problem of working out of the program application realising with the help ontology of the set subject domain the decision of a problem of processing and use of computable values.

Одной из наиболее ярких идей развития Интернет является инициатива по созданию «интеллектуального» Интернета, получившая название "Semantic Web". Главная задача здесь состоит в разработке и внедрении ряда новых подходов, которые позволили бы в цепочку «Интернет – пользователь» ввести ещё одно звено – «автоматический агент», который был бы наделён достаточным интеллектом, чтобы самостоятельно выполнять поиск и обобщение необходимой пользователю информации. Однако, для того, чтобы автономные программные модули были в состоянии без помощи человека обрабатывать информацию в Интернет, необходимо сделать информацию, содержащуюся в сети, понятной не только человеку, но и машине.

Одной из ключевых задач Semantic Web является нахождение семантического подобия между объектами в Internet. В качестве таких объектов могут выступать как Web-ресурсы, так и отдельные понятия – составляющие Web-контента. Вопрос подобия является предметом широких исследований в компьютерных науках, искусственном интеллекте, психологии, и лингвистике. В частности, извлечение информации имеет давние традиции поиска меры подобия между документами.

Представление знаний в рамках Semantic Web осуществляется с помощью механизма онтологий. Основной целью использования онтологий является обмен данных не только на общем синтаксическом, но и на распределенном семантическом уровне. Наиболее часто используются распределенные онтологии. Это обусловлено тем, что в средах с множеством информационных систем независимые системы могут иметь собственные, предназначенные для них модели и, соответственно, собственные онтологии.

Онтологии рассматриваются как решение проблемы разнородности данных в Web. Однако онтологии сами могут являться причиной разнородности. Концепты в онтологиях представлены словами на естественном языке. В то же время значения и понимание слов являются отличными в разных обществах. В случае наличия двух онтологий одному и тому же понятию могут даваться разные имена, или эти понятия могут быть определены по-разному, например, обе онтологии могут выражать одни и те же знания на разных языках.

С увеличением доступа к разнородным и независимым хранилищам данных определение семантического подобия или различия понятий онтологий является ключевым при извлечении и интеграции информации.

Применяемые сейчас технологии обработки текстовой документации статистическими методами принципиально не в состоянии обеспечить высокой (близкой к единице) точности поиска информации из-за отсутствия контекстного понимания смысла документа. Разрабатываемые же методы консорциума W3C в состоянии учесть столь важную составляющую как контекст при обработке документов. На данный момент разработаны и утверждены стандарты Unicode, URI, XML+XML Schema, RDF +RDF Schema и OWL (был утверждён как W3C рекомендация). Но перечисленные стандарты описывают только способ представления информации, что же касается программных модулей для обеспечения доступа к компонентам стандартизированных файлов, то они разработаны только для Unicode, URI, XML+XML Schema, RDF +RDF Schema. Для OWL

существует несколько тестовых библиотек для обеспечения доступа, но они еще не в полной мере поддерживают набор необходимых функций. Но даже разработанные компоненты для RDF, RDFS, OWL уже позволяют начать работу над созданием принципиально отличных, значительно превосходящих по степени интеллектуализации информационных систем. Предложенный набор технологий уже позволяет приступить к решению сложных задач, связанных с повторным использованием информации, интеграцией информации в рамках единого информационного пространства и проведением интеллектуального вывода.

Практическое применение такие системы имели бы в широком классе научно производственных комплексов:

- в системах менеджмента знаний больших предприятий и консорциумов;
- в системах поддержки сравнительного анализа украинских и европейских стандартов высшего образования по направлениям подготовки со сравнением оценки украинских и европейских кредитов;
- для поддержки высокоинтеллектуальных мультимедийных систем, использующих средства искусственного интеллекта;
- для поддержки систем интеллектуализации компьютерных интерфейсов;
- в системах интеграции баз знаний;
- поддержка распределённых систем, основанных на применении интеллектуальной мультиагентной технологии.

В последние годы разработка онтологий – формальных явных описаний терминов предметной области и отношений между ними – переходит из мира лабораторий по искусственному интеллекту на рабочие столы экспертов по предметным областям. Во всемирной паутине онтологии стали обычным явлением. Онтологии в сети варьируются от больших таксономий, категоризирующих веб-сайты (как на сайте Yahoo!), до категоризаций продаваемых товаров и их характеристик (как на сайте Amazon.com). Консорциум WWW (W3C) разрабатывает RDF (Resource Description Framework) (Brickley and Guha 1999), язык кодирования знаний на веб-страницах, для того, чтобы сделать их понятными для электронных агентов, которые осуществляют поиск информации.

Онтология определяет общий словарь для ученых, которым нужно совместно использовать информацию в предметной области. Она включает машинно-интерпретируемые формулировки основных понятий предметной области и отношения между ними.

Почему возникает потребность в разработке онтологии? Вот некоторые причины:

- Для совместного использования людьми или программными агентами общего понимания структуры информации.
- Для возможности повторного использования знаний в предметной области.
- Для того чтобы сделать допущения в предметной области явными.
- Для отделения знаний в предметной области от оперативных знаний.
- Для анализа знаний в предметной области.

Совместное использование людьми или программными агентами общего понимания структуры информации является одной из наиболее общих целей разработки. К примеру, пусть, несколько различных веб-сайтов содержат информацию по медицине или предоставляют информацию о платных медицинских услугах, оплачиваемых через Интернет. Если эти веб-сайты совместно используют и публикуют одну и ту же базовую онтологию терминов, которыми они все пользуются, то компьютерные агенты могут извлекать информацию из этих различных сайтов и накапливать ее. Агенты могут использовать накопленную информацию для ответов на запросы пользователей или как входные данные для других приложений.

Обеспечение возможности использования знаний предметной области стало одной из движущих сил недавнего всплеска в изучении онтологий. Например, для моделей многих различных предметных областей необходимо сформулировать понятие времени. Это представление включает понятие временных интервалов, моментов

времени, относительных мер времени и т.д. Если одна группа ученых детально разработает такую онтологию, то другие могут просто повторно использовать ее в своих предметных областях. Кроме того, если нам нужно создать большую онтологию, мы можем интегрировать несколько существующих онтологий, описывающих части большой предметной области. Мы также можем повторно использовать основную онтологию, такую как UNSPSC, и расширить ее для описания интересующей нас предметной области.

В литературе по искусственному интеллекту содержится много определений понятия онтологии, многие из которых противоречат друг другу. Будем считать, что онтология – формальное явное описание понятий в рассматриваемой предметной области (классов (иногда их называют понятиями)), свойств каждого понятия, описывающих различные своства и атрибуты понятия (слотов (иногда их называют ролями или свойствами)), и ограничений, наложенных на слоты (фацетов (иногда их называют ограничениями ролей)). Онтология вместе с набором индивидуальных экземпляров классов образует базу знаний. В действительности, трудно определить, где кончается онтология и где начинается база знаний.

Для решения выше изложеной задачи, предлагаю использовать языки описания WEB-страниц RDF и OWL онтологии для представления предметной области различного рода информационных WEB-ресурсов - учебных, научных, познавательных, технических и проч.

В предлагаемой технологии выделяем следующие этапы:

- Разработка структуры WEB-ресурса в виде онтологии: создание иерархии классов, свойств, отношений между классами, их характеристик для описания WEB-ресурса. Затем здесь же создаем соответствующую базу знаний, в которой могут быть определены такие понятия, как страница, раздел, меню, параграф и др.
- Программная реализация преобразования онтологий в другой, более распространенный, формат.

Для работы данной технологии, требуется некоторая среда, которая бы связывала разработанные базы знаний и шаблоны в единый работающий WEB-ресурс. Будем вести разработку подобной среды на основе программных технологий Java Servlets. Архитектура нашего WEB-ресурса представлена на рисунке 1.

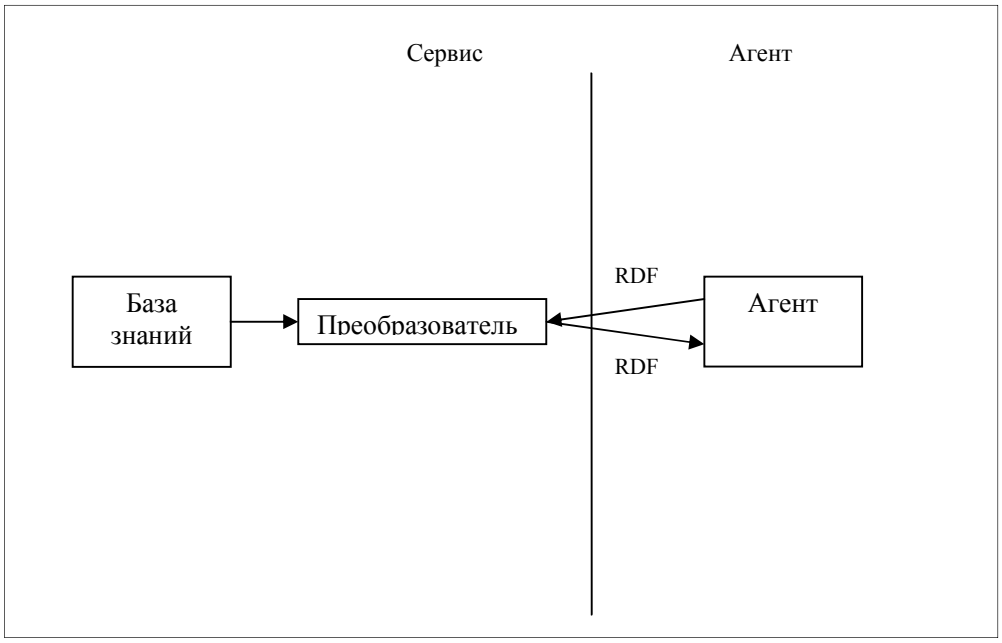

Рисунок 1 – Архитектура WEB-ресурса

Страница состоит из контейнеров и заголовка. В свою очередь, контейнеры могут включать составные (Parts), представляют собой логические компоненты любого WEB-ресурса, например, изображения или текста. Некоторые из составных имеют специфические свойства, например, у изображений есть свойство, указывающее на имя файла. Все составные могут являться ссылками на другие WEB-ресурсы. Контейнеры типа Content, кроме того, включат в себя элементы информационного наполнения сайта. Структура WEB-страницы представлена на рисунке 2.

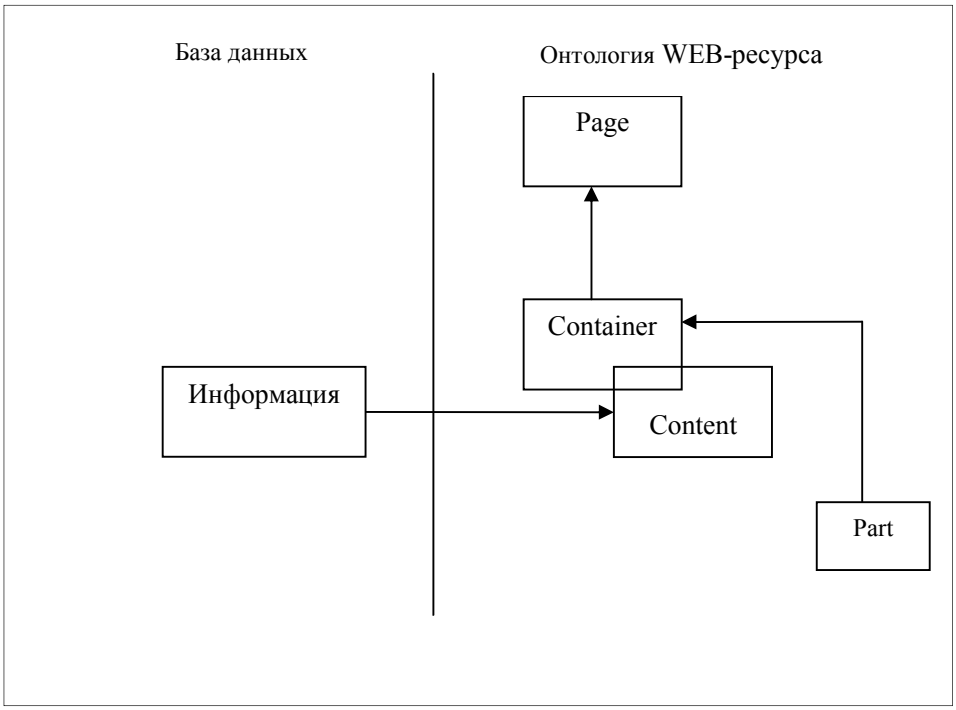

Рисунок 2 – Структура WEB-страницы.

Аархитектура онтологии контента является достаточно гибким решением, позволяющим максимально использовать полученные ресурсы системы в сети интернет, и позволит агентам использовать большее количество инструментов для получения запрашиваемой информации. Подобное решение на данный момент является наиболее эффективным, т.к. дает возможность взаимодействия системы и с пользователями и с другими системами, без использования дополнительных узкоспециализированных средств.

Кроме того, все использованные технологии являются бесплатными и свободно распространяемыми в сети интернет, причем большинство уже поддерживается распространенными пользовательскими программами.

К достоинствам представленной архитектуры можно отнести легкую повторную используемость созданных онтологий, универсальность, интероперабельность информации.

К недостаткам можно отнести сложность разработки онтологий, уникальность подобной архитектуры и, следовательно, отсутствие существующих стандартов.