УДК 002.53:004.89[А.Ф.1]

ФОРМИРОВАНИЕ СЕМАНТИЧЕСКИХ МЕТАДАННЫХ ДЛЯ ОБЪЕКТОВ СИСТЕМЫ УПРАВЛЕНИЯ ЗНАНИЯМИ

А.Ф. Тузовский

Институт «Кибернетический центр» ТПУ
Томский научный центр СО РАН
E-mail: TuzovskyAF@kms.cctpu.edu.ru

Предлагаются методы формирования семантических метаданных для различных объектов системы управления знаниями. Рассмотрен метод ручного аннотирования различных объектов с использованием редактора метаданных. Для документов предложен метод полуавтоматического аннотирования с использованием поверхностного лингвистического анализа.

Введение

В соответствии с онтолого-семантическим подходом к созданию систем управления знаниями (СУЗ) организации [1] все объекты (документы, специалисты, подразделения, базы данных и т. п.), содержащие знания, описываются с использованием различных метаданных. Метаданные – это данные, описывающее контекст (*context* – от лат. *связь*) и контент (от англ. *content* – содержание) объектов. Контекстные метаданные описывают связь объекта с другими объектами системы, а контентные метаданные описывают содержимое объекта (т. е., имеющиеся в объекте знания). Использование метаданных, в особенности контентных (семантических), позволяет эффективно решать такие задачи работы со знаниями как поиск, категоризация и рекомендация знаний. Аннотирование – процесс создания метаданных – может выполняться как с участием человека (там, где это необходимо), так и без его участия – автоматически. Однако в связи с тем, что задача понимания текстов на естественном языке до сих пор в полной мере не решена, не представляется возможным составление качественных контентных метаданных без участия человека. В лучшем случае, этот процесс является полуавтоматическим, когда программы предлагают варианты утверждений для контентных метаданных, а человек анализирует их и либо принимает, либо редактирует или отвергает.

Анализ состава информации современной организации показывает, что основная ее часть содержится в виде текстов на естественном языке – более 80 %, в бумажной и электронной формах. В связи с этим, одной из наиболее сложных задач в построении СУЗ является разработка методов составления достаточно точных контентных метаданных для текстовых документов.

Онтологический подход к решению задачи аннотирования документов

Онтологический подход предполагает в качестве содержимого метаданных использование элементов онтологии [2]. Контентные (семантические) метаданные $M_c=\{s_1, s_2, ..., s_m\}$ это наборы семантических утверждений (триплетов) s_i, которые имеют вид $s_i=(c, r, o, v)$, где c – это субъект утверждения (понятие, или экземпляр – контекстные метаданные некоторого понятия), o – объект (экземпляр – контекстные метаданные некоторого понятия), r – отношение между субъектом и объектом, а v – весовой коэффициент, который оценивает значимость данного утверждения для описания объекта знаний. При этом понятия и отношения должны быть описаны в онтологии O, а экземпляры описываются контекстными метаданными онтологической базы знаний. Без использования весовых коэффициентов примерами утверждений являются следующие триады $<C, R, C>$, $<I, R, I>$, $<C, R, I>$, $<I, R, V>$, $<C, R, NULL>$, $<C, NULL, NULL>$, $<I, R, NULL>$, $<I, NULL, NULL>$, где C – понятие; I – экземпляр понятия; R – связь; A – атрибут; V – значение атрибута (текстовое или числовое).

Задача аннотирования заключается в создания семантических метаданных, т. е. в формировании множества утверждений (триплетов), на основе некоторой онтологии и соответствующей ей базе знаний. Возможны ручной и полуавтоматический варианты решения данной задачи.

Ручной вариант реализации заключается в создании редактора метаданных, который позволяет пользователю, с помощью специального интерфейса выбрать элементы утверждений используя онтологию некоторой предметной области и свои знания об аннотируемом объекте (документе, специалисте и т. п.). Основной задачей интерфейса является предоставление возможности конструирования метаданных с одновременной навигацией по онтологии, в том числе и с интерактивной визуализацией отдельных ее частей.

Полуавтоматический вариант реализации предполагает создание подсистемы, которая анализирует объект знаний, имеющий текстовое содержание, а после этого предоставляет пользователю «начальный вариант» семантического метаописания, которые пользователь может отредактировать. При этом экономится время специалиста на ознакомление с содержанием объекта.

Семантические метаданные применяются для описания объектов системы управления знаниями [2] и используются в процедурах семантической обработки информации. Объекты могут либо иметь, либо не иметь текстовое описание. В зави-

симости от этого формирование семантических метаданных будет выполняться различными способами. В данном исследовании разработан метод формирования семантических метаданных, который определяет правила выбора предикатов и объектов из онтологии, а также определяет алгоритм поиска понятий и экземпляров в тексте.

Формирование семантических метаданных объекта знаний должен выполнять человек. Он должен в соответствии с сущностью предмета описания определять элементы семантических метаданных. Элементы представляют собой либо триплеты со структурой «субъект – предикат – объект», либо отдельные понятия или экземпляры из онтологии, которые будем называть «субъект». Создавая элемент семантических метаданных, человек обязательно должен указать «субъект». После этого он может дополнительно указать «предикат» и «объект».

Если субъект указывается человеком таким образом, чтобы отражать сущность предмета описания, то на выбор предиката и объекта накладываются дополнительные ограничения, которые вытекают из правил формирования высказываний дескриптивной логики.

Множество возможных предикатов в триплете ограничивается выбранным субъектом триплета. После выбора предиката человек должен обязательно указать объект триплета. Множество возможных объектов зависит от выбранного предиката. Возможные значения предиката определяются либо областью конкретных значений атрибута, либо областью значений отношения.

Если семантические метаданные формируются на основании текстового описания объекта, то в дополнение к правилам выбора предикатов и объектов используется *алгоритмом поиска понятий и экземпляров в тексте*. Это позволяет частично автоматизировать процесс выбора субъекта из онтологии. С этой целью текстовое описание анализируется на наличие понятий и экземпляров, которые могут выступать в качестве субъектов в элементах семантических метаданных.

Человек, формирующий семантические метаданные, должен отредактировать полученное множество понятий и экземпляров:

- удалить элементы, не отражающие сущность объекта описания;
- устранить многозначность, если множество содержит элементы с одинаковыми лексическими метками;
- дополнить множество понятиями и экземплярами, не найденными алгоритмом.

После этого элементы множества могут быть использованы для формирования триплетов в соответствии с описанными выше правилами выбора предикатов и объектов.

При работе семантического портала [3] рассмотренный метод используется для формирования семантических метаданных объектов различного типа. Например, в процессе семантического описания знаний специалиста не используется алгоритм поиска понятий и экземпляров в тексте, так как нет соответствующего текстового описания его знаний. Для документов семантические метаданные создаются на основании их текстового содержания, что позволяет задействовать алгоритм поиска понятий и экземпляров.

Для составления семантических метаданных был разработан набор программ для выполнения ручного и полуавтоматического аннотирования: редактор контентных метаданных для ручного аннотирования объектов и компонент полуавтоматического аннотирования.

Реализация ручного семантического аннотирования документов

Выполнение ручного семантического аннотирования объектов в составе семантического портала реализуется с помощью редактора контентных (семантических) метаданных. Данный редактор состоит из двух частей: компонента поддержки интерфейса (*WebControl*) и связанной с ним HTML страницы, которая также включает поставщика данных (*DataSource*) и компонента визуализации структуры онтологии (Навигатор онтологии).

Компонент поддержки интерфейса отвечает за визуализацию редактора, выполнение логики работы редактора и реализован на языке *JavaScript*. Программа на *JavaScript* используется как для управления так и для разбора данных от источника данных *DataSource*. Поставщик данных *DataSource*, получив данные (описание онтологии), передает их в формате *XML* скрипту, выполняющему поддержку интерфейса. Типичные запросами к источнику данных являются: получение всех понятий и экземпляров онтологии по некоторой лексической метке; получение всех возможных свойств (отношений и атрибутов) для заданного понятия или экземпляра; получение имени домена для отношения (домен может быть сложным). Схема работы компонента «Навигатор онтологии» показана на рис. 1.

Элемент *WebControl* отвечает за загрузку *JavaApplet*, передачи ему вызовов а также предоставление программного интерфейса к *JavaApplet*. Элемент *JavaApplet* отвечает за функции рисования и решение задач топологии. Он посылает запрос к своему источнику данных, принимает XML данные, разбирает их (используется грамматический анализатор (*парсер*) *Nano* – для мобильных устройств и телефонов, выбран из-за малого размера) и визуализирует данные, имеет функцию отображения дерева значительной сложности, однако целесообразно отображать не более 3–4 уровней.

На данном этапе источник данных генерирует вложенность первого уровня для экономии места на экране (рис. 2).

Поставщик данных *DataSource* получив запрос от *JavaApplet* данной страницы, выполняет обраще-

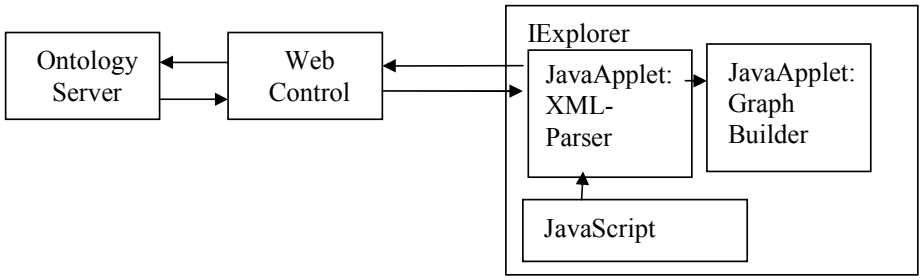

Рис. 1. *Взаимодействие элементов компонента «Навигатор онтологии»*

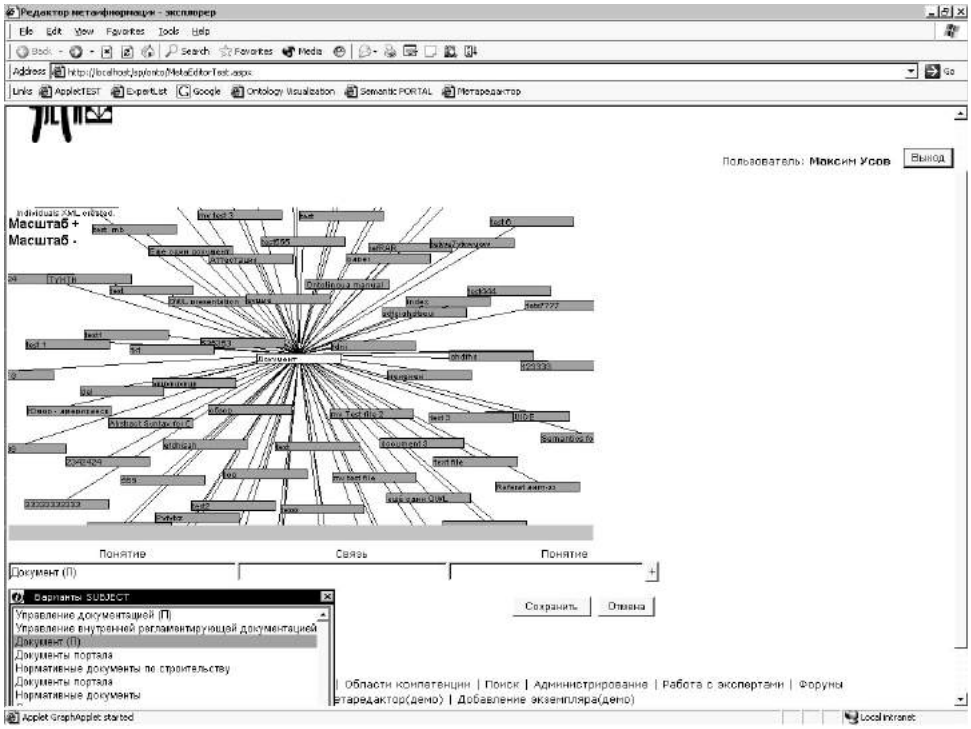

Рис. 2. *Интерфейс редактора семантических метаданных: отображение экземпляров выбранного понятия*

ние к онтологии и формирует XML-описание для апплета, в котором содержится информация об иерархической структуре онтологии.

Интерфейс пользователя состоит из строк, включающих по три поля. В первом поле указывается выбранное понятие онтологии (подлежащее), во втором – выбранное отношение (сказуемое) и в третьем заданное значение. Подобных строк может быть неограниченное количество. Редактор снабжен оперативной подсказкой, которая активируется после того, как пользователь перестал изменять текст в течении некоторого времени.

Полуавтоматическое семантическое аннотирование документов

В рамках проводимых исследований по полуавтоматическому аннотированию, был разработан подход, базирующийся на использовании лингвистических методов морфологического, синтаксического и поверхностного семантического (общеописательного) анализов текста документов на естественном языке [4]. Результатом поверхностного семантического анализа предложения на естественном языке является связанный граф (в некоторых случаях множество графов), который дополняет предложение набором связей из фиксированного словаря, являющегося, в свою очередь, прототипом метаинформации для предложения. Использование в алгоритме анализа текста в качестве исходных данных результата поверхностного семантического анализа, позволит избавиться от множества неоднозначностей интерпретации, связанных с лингвистическими особенностями языка.

Следующим шагом к составлению семантической информации является этап сопоставления частей семантического графа элементам онтологии, фактически, здесь происходит проецирование одного графа на другой. Проецирование возможно только при наличии соответствующих правил. Эти правила должны представляют собой некоторый набор лингвистических шаблонов.

Конечной задачей анализа текста является выявление поверхностного (абстрактного) смысла, выраженного в виде набора утверждений. Для его выполнения используется многоуровневая обработка текста, при этом входными данными каждого последующего уровня являются выходные данные предыдущего.

Основными уровнями обработки являются: *Графематический анализ* размечает текст на предложения, слова, цифры, буквы и названия; *Морфологический анализ* выявляет тип часть речи каждого слова, форму, склонение, число и т. д.; *Синтаксический анализ* определяет синтаксические связи внутри предложения между его частями, выполняет развернутый синтаксический разбор и *Поверхностный семантический анализ* по своей сути является измененным синтаксическим анализом.

В отличие от синтаксического анализа, где по фиксированным алгоритмам строятся связи на основании морфологических данных, здесь используют множество эвристик и данных о шаблонах построения предложения в языке. Результатом являются установленные абстрактные связи между частями. Нужно отметить, что данный вид связей по своей сути является уже частью онтологии верхнего уровня. Однако на этом этапе самого понятия онтологии нет, и модель предметной области не используется, вместо неё имеется множество правил и сведений о традициях использования языка (база примеров).

Данный вид анализа текста поддерживает одна из российских разработок «RML 2006» фирмы АОТ [5]. RML является продуктом с открытым кодом. Его модули реализованы в виде динамической библиотеки и поддерживают компонентную технологию ActiveX, что позволяет достаточно просто подключить библиотеку к приложению в среде .Net. Система АОТ имеет качественный синтаксический анализатор, который можно использовать для процесса генерации семантических метаданных.

Метод анализа текста

Подготовка. Вначале текст разбивается на предложения. Предложения передаются на вход синтаксического анализатора АОТ. Конечные и промежуточные результаты разбора заносятся в хэш-таблицу. Это делается для ускорения работы основного алгоритма, который нуждается в многократной нормализации и анализе каждого слова в предложении и предложения целиком. После этого входной текст нормализуется (приводится к начальным формам). Аналогичным образом обрабатывается и онтология – лексические метки всех понятий также нормализуются и заносятся в хэш-таблицу.

Выделение терминов. Нормализованные лексические метки понятий ищутся среди нормализованных последовательностей слов текста. Найденные или похожие термины (в том числе составные) выделяются и заносятся в список триплетов. Каждому найденному термину соответствует не полный триплет $<C, , >$, где C – это найденное понятие онтологии.

Нахождение экземпляров. Данный этап состоит из двух частей. Первая часть – аналогична фазе выделения терминов, но в этом случае ищутся только экземпляры понятий, которые уже находятся в онтологии, и каждому найденному экземпляру ставится в соответствие неполный триплет $<I, , >$, где I – это найденный экземпляр понятия онтологии. Вторая часть основана на наборе эвристических правил, которые выполняют предварительное выделение претендентов на право быть экземплярами. После этого начинается проверка дополнительных правил, например, если претендент:

- стоит рядом с понятием предметной области, то он является экземпляром и имеется высокая вероятность того, что это экземпляр именно этого понятия;
- начинается с большой буквы или записан в кавычках, то он также однозначно является экземпляром какого-то понятия.

Реализации системы полуавтоматического аннотирования

Основой созданной системы полуавтоматического составления семантических метаданных является компонент анализа текста [4]. Данный компонент реализован на основе платформы Microsoft Framework 2.0 с использованием технологии «Code-behind» применительно к документам, создаваемым с использование текстового редактора Microsoft Word 2003. Такой подход предполагает, что документы организации создаются на основе специально разработанного шаблона документов. Этот шаблон включает ссылку на исполняемый код компонента, выполняющего анализа текста, и представленного в виде динамической библиотеки, которая загружается на стороне клиента вместе с шаблоном документа. Таким образом, имеется возможность расширить функции обычного офисного средства (в данном случае – это Microsoft Word) до возможности полуавтоматического составления семантических метаданных на основе анализа текста и работы с онтологией.

Использование компонента, связанного с шаблоном документа, позволяет оперативно выполнить семантический анализ непосредственно из редактора текстов Microsoft Word и в результате получить семантические метаданные в виде триплетов, которые могут быть отредактированы и сохранены на сервере. Порядок взаимодействия программы построения метаданных с системой управления знаниями показан на рис. 3.

Связь с Web-сервером семантического портала системы управления знаниями осуществляется через Web-сервисы по протоколу SOAP, основанному на XML. Web-сервер имеет доступ к онтологии через сервер онтологий и к хранилищу метаданных, где хранятся метаданные для всех объектов портала, а также информация об их принадлежности к объектам.

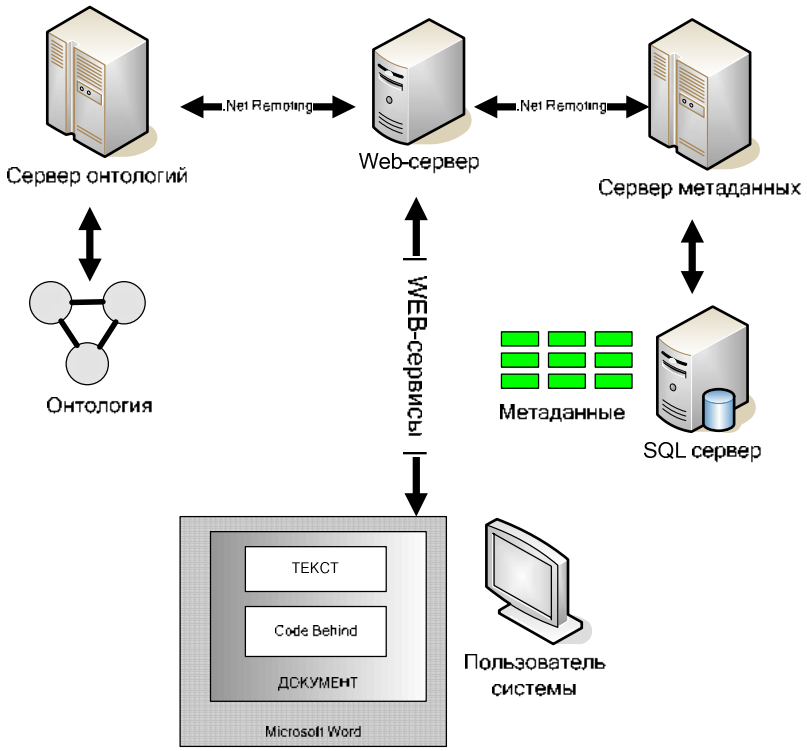

Рис. 3. *Порядок взаимодействия программы построения метаданных с системой управления знаниями*

СПИСОК ЛИТЕРАТУРЫ

1. Davenport T., Prusak L. Working Knowledge: how organizations manage what they know. – Boston: Harvard Business School Press, 1998. – 200 p.
2. Тузовский А.Ф. Разработка системы управления знаниями на основе единой онтологической модели // Известия Томского политехнического университета. – 2007. – Т. 310. – № 2. – С. 182–185.
3. Тузовский А.Ф. Архитектура семантического Web-портала // Известия Томского политехнического университета. – 2006. – Т. 309. – № 7. – С. 142–145.
4. Тузовский А.Ф., Усов М.В. Семантическое аннотирование документов // Информационные и системные технологии в индустрии, образовании и науке: Научные труды Междунар. симп. – Караганда, 2006. – С. 240–242.
5. Пакет документации к программному продукту RML // [Электронный ресурс]. – 2006. – Режим доступа: [http://www.aot.ru/technology.html].

Поступила 21.11.2006 г.