Известия Челябинского научного центра, вып. 3 (16), 2002

ИНФОРМАТИКА И ИНФОРМАЦИОННЫЕ СИСТЕМЫ

УДК 681 .3.068 + 658.512–52

МОДЕЛИРОВАНИЕ СЕМАНТИКИ И ПРАГМАТИКИ ДОКУМЕНТА В НОТАЦИИ ЯЗЫКА XML

С.Е. Коровин (1), А.В. Мельников, И.Л. Кафтанников (2)
e–mail: sergey_korovin@nm.ru (1), kil@comp.susu.ac.ru (2)

Южно–Уральский государственный университет, г. Челябинск, Россия

Статья поступила 19 сентября 2002 г.

На сегодняшний день существует два основных подхода к решению проблемы автоматизации семантического анализа естественно–языковых текстов.

Первый подход заключается в поиске методов интерпретации синтаксических и поверхностно–семантических конструкций естественного языка — ассоциации лексем и словокомплексов с некоторой соответствующей им системой понятий. Такая постановка проблемы семантического анализа позволяет достаточно эффективно решать задачи, непосредственно связанные со знаковой системой языка, — задачи семантического поиска, классификации, автоматического реферирования и т. п. Однако, поверхностно–семантические модели (словарь понятий, на который отражаются лексемы и словокомплексы, и правила этого отображения) — сложны, объемны и существенно изменяются от одной предметной области к другой, что значительно снижает эффективность их практического использования.

Вторым подходом к решению проблем семантического анализа является создание искусственной семиотической системы — семантической модели (глубинно–семантической модели). Семантическая модель представляет собой необъемную систему однозначных и строго структурированных понятий, полученных путем обобщения концептов (понятий) естественного языка. С семантической моделью ассоциируется формализованная нотация, еще более упрощающая автоматизацию анализа модели.

В последние несколько лет второй подход получает все большое распространение, поскольку уровень развития первого подхода не позволяет решать многие практические задачи. Одним из примеров, подтверждающих данную тенденцию, является развитие стека протоколов World Wide Web, базирующегося на расширенном языке разметки — XML [1]. Одной из задач данного стека является создание основы для автоматизации семантической обработки информации в Web. Эту задачу призвана решить связка XML + RDF, в которой язык XML играет роль синтаксиса (формализованной нотации), а модель RDF [2] — роль семантической модели (структурированного набора смысловых понятий). Узким местом данной связки (а возможно и всего стека протоколов) является модель RDF, представляющая собой, фактически, семантическую сеть (основным недостатком практически любой семантической сети является ее статичность, которая резко усложняет моделирование динамических явлений).

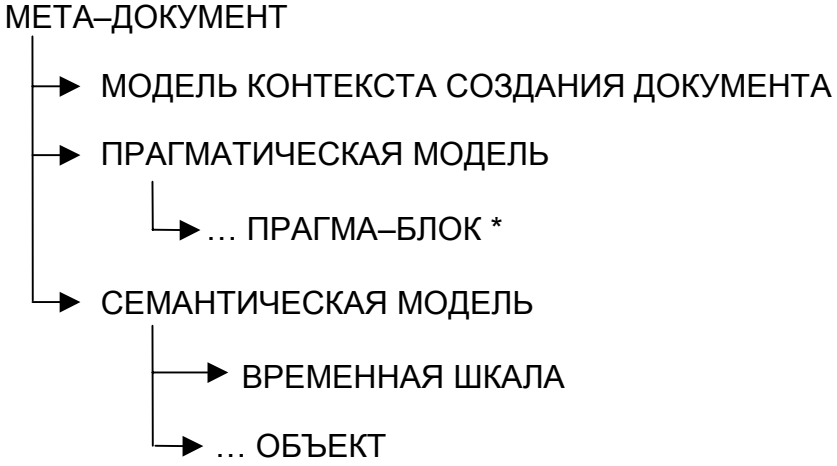
Создание семантических моделей, которые бы расширили возможности RDF, вызвало к жизни новое исследовательское направление, поддерживаемое WWW–консорциумом, — Semantic Web [4]. В его контексте созданы такие модели, как, например, DAML + OIL [3].

Описываемая в данной статье семантико–прагматическая модель документа (СПМД), так же, относится к направлению Semantic Web (использует нотацию языка XML) и расширяет функциональность существующих в рамках этого направления моделей, поддерживая представление динамических явлений и прагматики документов.

Семантико–прагматическая модель документа

Структура предлагаемой семантико–прагматической модели (будем, так же, называть ее мета–документом) состоит из трех базовых моделей (рисунок): модели контекста создания документа (описания автора, адресата и свойств документа), прагматической модели (описания прагматических блоков, составляющих документ) и семантической модели (динамического описания глубинной семантики документа).

Прагматическая модель базируется на идеи о соответствии целей автора документа в отношении адресата типам поверхностно–семантической организации текста (типам иллокутивных актов). Она представляет собой совокупность прагматических блоков, каждый из которых ставит в соответствие определенному участку текста цель автора и средство ее достижения — особый способ поверхностно–семантической организации этого участка текста. Например, цели «изменение волевого состояния адресата» могут соответствовать следующие средства: утверждение, предположение.



Структура семантико–прагматической модели документа

* «→…» означает, что элемент может присутствовать 0, 1 и больше раз.

В основе *семантической модели* лежит следующая идея: человек представляет окружающий его мир в виде объектов, характеризуемых свойствами и отношениями (взаимодействиями) между собой или, другими словами, — своими состояниями. Состояния объектов постоянно изменяются. Существует два способа рассмотрения этих изменений: временной (изменения состояний объектов рассматриваются относительно некоторого эталонного изменения — времени) и причинно–следственный (изменения состояний объектов рассматриваются относительно состояний объектов, с которыми они взаимодействуют). Таким образом, *динамическая* семантическая модель должна представлять собой описание временной и причинно–следственной составляющих изменения свойств и отношений взаимодействующих между собой объектов.

Движение объекта можно описывать двумя способами. Первый способ заключается в разбиении процесса изменения свойств и отношений объекта на статические состояния («мгновенные снимки»). Ему, в частности, соответствует математическая модель — абстрактный автомат. Второй способ описания заключается в формировании последовательности переходов (т. е. единичных изменений) объекта. Он позволяет более точно и компактно описывать каждое конкретное изменение и по своей сущности напоминает табличное представление функции. СПМД поддерживает оба этих способа (и в этом смысле, напоминает по своей структуре сети Петри, хотя понятие перехода здесь существенно отличается от понятия перехода в сетях Петри).

Как видно из рисунка, элементами верхнего уровня семантической модели являются «временная шкала» и «объект». Объект является ключевым элементом семантической модели. При моделировании явления, описываемого в документе, в явлении выделяются взаимодействующие объекты, после чего движение каждого объекта описывается отдельно.

Временная шкала разбивает моделируемое явление на несколько последовательных интервалов, путем описания ключевых моментов. Описание осуществляется либо посредством указания абсолютного времени момента, либо путем его ассоциации с некоторым ключевым событием. Заданные таким образом моменты используются далее при описании временных промежутков конкретных переходов и состояний объектов.

Объект характеризуется своими свойствами; отношениями (какие роли он играет в этих отношениях, с какими объектами он ими связан, каков тип данного отношения — классификационный, ролевой); и, если данный объект является системой, — структурными связями (парами вида «объект 1 — объект 2», множество которых позволяет задать структуру системы).

При описании объекта, прежде всего, задаются статические свойства, отношения и структурные связи (те, которые не изменяются на всем протяжении моделируемого явления; например, наименование объекта). Они размещаются внутри элемента «объект» и не входят в элементы «переход» и «состояние».

Далее осуществляется описание движения объекта. Для этого вводится последовательность переходов (они группируются друг за другом в порядке их возникновения). Каждый переход содержит в себе следующие элементы: «характер», «время», «условие», «причина», «следствие» и набор элементов, которые, собственно, составляют содержание перехода («свойство», «отношение», «структурная связь»).

Характер описывает сущность изменения: появление, прекращение, изменение, совершение (появление или исчезновение свойства, отношения; изменения значения свойства или роли отношения, совершения действия и т. п.).

Время ассоциирует данный переход с одним из интервалов временной шкалы. Для этого в него входят такие элементы, как «В_МОМЕНТ», «ДО» и «ПОСЛЕ», значения которых — моменты абсолютного времени или ключевые события. Множество таких элементов определяет конкретный временной интервал данного перехода.

Условие, причина и следствие характеризуют данный переход, как элемент некоторой причинно–следственной связи. Эти элементы указывают на переходы, отношения, конкретные элементы переходов (свойства, отношения, структурные связи), которые являются соответственно условиями, причинами и следствиями данного перехода.

Помимо описания движения в виде переходов, модель, так же, поддерживает описание движения в виде совокупности *состояний*, расположенных в порядке их смены. Этот уровень описания является более абстрактным, чем основной способ и дополняет его. Он присутствует в модели как минимум в виде пары: начальное и конечное состояния. Однако эксперт, формирующий модель, может ввести в нее, так же, любое число промежуточных состояний. Каждое состояние содержит описание временного промежутка, в течении которого оно имеет смысл, и всех свойств, всех отношений и всех структурных связей объекта, которыми он обладает в данном временном промежутке.

Рассмотрим пример моделирования естественно–языкового сообщения с использованием предложенной СПМД. Пусть имеется следующая ситуация. Газета «Мир кино» сообщила: «*Появились слухи о том, что Владимир Машков подписал контракт на участие в Голливудском проекте. Доподлинно известно, что 12 марта он вылетел из Москвы в Лос–Анджелес*». Тогда семантико–прагматическое представление данного сообщения выглядит в нотации XML так:

```
<?xml version="1.0" encoding="cp866"?>
<МЕТА–ДОКУМЕНТ>
<ДОКУМЕНТ>
        <НАЗВАНИЕ>Сообщение газеты “Мир Кино”</НАЗВАНИЕ>
        <ДАТА>20.03.2000</ДАТА>
        <УЧАСТНИК_ОБЩЕНИЯ Имя=”Мир Кино” Тип=”Автор”>
          <СОЦИАЛЬНЫЙ_СТАТУС>
                <ХАРАКТЕРИСТИКА>Печатное издание</ХАРАКТЕРИСТИКА>
          </СОЦИАЛЬНЫЙ_СТАТУС>
        </УЧАСТНИК_ОБЩЕНИЯ>
</ДОКУМЕНТ>
```

```
<ПРАГМАТИЧЕСКАЯ_МОДЕЛЬ>
   <ПРАГМА–БЛОК Имя=”Ходят слухи”>
        <ЦЕЛЬ Тип=”Передача_информации”/> <ЦЕЛЬ Тип=”Изменение_эмоц_состония”/>
        <СРЕДСТВО Тип=”Предположение”/>
        <ТЕКСТ> Появились слухи о том, что Владимир Машков подписал
                контракт на участие в Голливудском проекте.
        </ТЕКСТ>
    </ПРАГМА–БЛОК>

   <ПРАГМА–БЛОК Имя=”Доподлинно известно”>
        <ЦЕЛЬ Тип=”Передача_информации”/> <ЦЕЛЬ Тип=”Изменение_эмоц_состония”/>
        <СРЕДСТВО Тип=”Утверждение”/>
        <ТЕКСТ> Доподлинно известно, что 12 марта он вылетел из Москвы в Лос–Анджелес.
       </ТЕКСТ>
    </ПРАГМА–БЛОК>
</ПРАГМАТИЧЕСКАЯ_МОДЕЛЬ>

<СЕМАНТИЧЕСКАЯ_МОДЕЛЬ>

  <ОБЪЕКТ Имя=”Владимир Машков”>
    <НАЧАЛЬНОЕ_СОСТОЯНИЕ>
         <СВОЙСТВО> <ИМЯ>Место нахождения</ИМЯ> <ЗНАЧЕНИЕ>Москва</ЗНАЧЕНИЕ>
        </СВОЙСТВО>
    </НАЧАЛЬНОЕ_СОСТОЯНИЕ>

    <ПЕРЕХОД Характер=”Появление” UID=”Переход#1”>
         <ВРЕМЯ>  <ДО Измерение=”Время” Значение=”12.03.2000”/> </ВРЕМЯ>
         <ОТНОШЕНИЕ Имя =”Участие в проекте”  UID=”Отношение#1”>
            <РОЛЬ>Актер</РОЛЬ>
            <СВОЙСТВО><ИМЯ>Место проекта</ИМЯ> <ЗНАЧЕНИЕ>Голливуд</ЗНАЧЕНИЕ>
            </СВОЙСТВО>
         </ОТНОШЕНИЕ>
    </ПЕРЕХОД>

    <ПЕРЕХОД Характер=”Изменение” UID=”Переход#2”>
         <ВРЕМЯ> <В_МОМЕНТ Измерение=”Время” Значение=”12.03.2000”/> </ВРЕМЯ>
         <ПРИЧИНА> <ССЫЛКА UID=”Отношение#1”/> </ПРИЧИНА>
         <СВОЙСТВО UID=”Свойство#1–2”>
            <ИМЯ>Место нахождения</ИМЯ> <ЗНАЧЕНИЕ>Лос–Анджелес</ЗНАЧЕНИЕ>
        </СВОЙСТВО>
    </ПЕРЕХОД>

    <КОНЕЧНОЕ_СОСТОЯНИЕ>
        <СВОЙСТВО Ссылка=”Свойство#1–2”/> <ОТНОШЕНИЕ Ссылка=”Отношение#1”/>
    </КОНЕЧНОЕ_СОСТОЯНИЕ>

   </ОБЪЕКТ>

  </СЕМАНТИЧЕСКАЯ_МОДЕЛЬ>
</МЕТА–ДОКУМЕНТ>
```

Таким образом, предложенная семантико–прагматическая модель позволяет описывать на формальном языке (XML) аналитическую информацию о смысловом содержимом документа. Поскольку указанная информация представляется в формальном виде, то появляется возможность достаточно быстро создавать различные алгоритмы ее анализа. А поскольку модель использует для записи аналитической информации общепринятую нотацию (язык XML), имеется возможность использования стандартных программных инструментов (XML–анализаторов), которые значительно облегчают и ускоряют создание этих алгоритмов анализа.

В частности, в рамках данного исследования разработаны несколько типовых алгоритмов:

– Анализ отношений объекта. Этот анализ проводится в тех случаях, когда нужно охарактеризовать некоторый объект с точки зрения отношений, в которых он состоит (какие типы отношений наиболее характерны для этого объекта; какие он играет в них роли; какова динамика их изменения).
– Анализ окружения объекта. Этот анализ дает возможность узнать — с кем общался (состоял в отношениях, в отношении кого совершал действия) интересующий объект.
– Анализ истории изменений объекта. Этот анализ дает возможность проследить ретроспективу реальных изменений, произошедших с интересующим объектом за определенный период.

Заключение

Предлагаемая семантико–прагматическая модель документа:

– относится к перспективному научному направлению, поддерживаемому международной организацией WWW–консорциум, — Semantic Web;
– лежит в основе реализации сложных алгоритмов семантического анализа (в частности, анализа динамики отношений между определенной группой объектов; анализа истории изменений, происходящих с интересующим объектом; анализа динамики окружения некоторого объекта;
– поддерживает представление динамических явлений, которые, на данный момент, не рассматриваются другими семантическими моделями, входящими в направление Semantic Web (для моделирования динамических явлений в СПМД были включены специальные XML–конструкции, адаптирующие метод сетей Петри (метод представления динамики объекта в виде кортежа его состояний и переходов) к моделированию естественно–языкового описания динамики объекта);
– включает в себя прагматические элементы, что позволяет описывать помимо смыслового содержимого документа — цели автора и ассоциированные с ними языковые средства (типы иллокутивных актов), которые он использовал.

Список литературы

1. Extensible Mark–up Language (XML) 1.0. (Second Edition) // W3C Recommendation 6 October 2000. http://www.w3.org/TR/REC–xml
2. Resource Description Framework (RDF) Model and Syntax Specification // W3C Recommendation 22 February 1999. http://www.w3.org/TR/REC–rdf–syntax
3. Annotated DAML+OIL Ontology Markup // W3C Note 18 December 2001. http://www.w3.org/TR/daml+oil–walkthru
4. Semantic Web History: Nodes and Arcs 1989–1999 // The WWW Proposal and RDF. http://www.w3.org/1999/11/11–WWWProposal/