На правах рукописи

КОРОВИН Сергей Евгеньевич

ДИНАМИЧЕСКАЯ МОДЕЛЬ СЕМАНТИКИ И ПРАГМАТИКИ ДОКУМЕНТОВ НА БАЗЕ РАСШИРЕНИЯ ЯЗЫКА XML

Специальность 05.13.01 – "Системный анализ, управление и обработка информации (промышленность)"

Автореферат
диссертации на соискание ученой степени
кандидата технических наук

Челябинск

2002

Работа выполнена в Южно-Уральском государственном университете.

Научный руководитель – доктор технических наук, профессор Мельников А.В.

Официальные оппоненты:
доктор технических наук, профессор Казаринов Л.С.,
доктор технических наук, профессор Васильев В.И.

Ведущая организация – Челябинский филиал ОАО "Уралсвязьинформ".

Защита состоится 4 декабря 2002 года, в 14 часов, на заседании диссертационного совета Д 212.298.03 при Южно-Уральском государственном университете по адресу: 454080, г.Челябинск, пр. им. В.И. Ленина, 76, конференц-зал ЮУрГУ (ауд. 244).

С диссертацией можно ознакомиться в библиотеке Южно-Уральского государственного университета.

Автореферат разослан 29 октября 2002 г.

Ученый секретарь
диссертационного совета А.М. Коровин

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

В диссертации изложены основные научные результаты, полученные и опубликованные в 1999–2002 гг., связанные с разработкой семантико-прагматической модели документа, поддерживающей представление динамических явлений, описываемых в документах на естественном языке, а так же прагматики этих документов – в нотации XML (Extensible Markup Language). Данная модель расширяет ряд существующих алгоритмов семантического анализа, обеспечивая реализацию алгоритмов, связанных с исследованием динамики явлений, описываемых в документах, и алгоритмов анализа целей автора и средств их выражения. Практическую реализацию модели обеспечивает метод автоматического формирования семантико-прагматических представлений документов в процессе их синтеза. Метод основан не на лингвистическом анализе документов, а на применении шаблонов, создаваемых для различных классов документов на основе семантико-прагматической модели.

Концептуальные проблемы создания моделей и систем для интеллектуального анализа информации рассматриваются в научных трудах Г. Буча, П. Коуда, О.В. Логиновского, Д.А. Поспелова, В.М. Тарасова, Р.П. Чапцова, Э. Йордана и др.

Исследованиям проблемы построения семантических моделей естественно-языковых текстов посвящены работы Аверкина А.Н., Ю.Д. Апресяна, А.К. Жолковского, Ю.А, Загорулько, И. Катца, М.М. Мастермана, И.А. Мельчук, А.С. Нариньяни, Э.В. Попова, В.Ш. Рубашкина, Э.Ф. Скороходько, Г. Скрэгга, У.А. Уилкса, Ч. Филмора, И.А. Фодора, В.Ф. Хорошевского, Р.С, Шенка, С. Шапиро и др. Проблема построения прагматических моделей и их связи с семантическими моделями отражена в работах В.В, Богданова, Л. Бирнбаум, Дж. Мей, А.С. Нариньяни, Дж. Остина, Р.С, Шенка и др.

Проблемы построения семантических моделей неразрывно связаны с проблемой представления знаний (последняя является методологией построения семантических моделей). Представлению знаний посвящены работы Д.Г. Боброва, Т. Виноград, И.П. Кузнецова, М. Минского, Ч. Морриса, А.С. Нариньяни, Д.А. Поспелова, Э.В. Попова, Л.К. Шуберта и др.

В последние несколько лет большая концентрация усилий по созданию семантических моделей и языков представления знаний наблюдается вокруг технологии XML, развитие которой поддерживается международной организаций WWW-консорциум. В рамках технологии WWW было создано новое направление – Semantic Web, основная задача которого заключается в создании семантических моделей электронных документов, использующих в качестве своей нотации язык моделирования структур данных – XML. Эти модели должны лечь в основу решения задач семантического анализа интернет-документов, в частности, и документов – вообще. На данный момент среди семантических моделей, использующих нотацию XML, нет моделей, которые бы поддерживали представление динамически развивающихся (во времени и

в соответствии с причинно-следственными связями) процессов. Восполнение этого пробела – одна из основных задач предлагаемой в работе модели.

Актуальность темы. На сегодняшний день существует два основных подхода к решению проблем 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 поиска, классификации, автоматического реферирования и т.п. Однако, поверхностно-семантические модели (словарь понятий, на который отражаются лексемы и словокомплексы, и правила этого отображения) чрезвычайно сложны и меняются от одной предметной области к другой, что значительно снижает эффективность их практического использования (в частности, ограничивает разнообразие алгоритмов семантического анализа).

Вторым подходом решения 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 искусственная нотация, еще более упрощающая автоматизацию анализа модели. Поскольку гибкий и сложный синтаксис естественного языка заменяется искусственной нотацией, а семантика естественного языка – формализуется, данный подход потенциально позволяет реализовывать значительно более сложные алгоритмы семантического анализа

В последние несколько лет второй подход получает все большое распространение. Наиболее ярким подтверждением данной тенденции является быстрое развитие нового направления, поддерживаемого WWW-консорциумом, – Semantic Web. В его контексте создана семантическая модель RDF. Прелагаемая в данной диссертационной работе семантико-прагматическая модель документа является ее альтернативой и дополняет ее, поддерживая представление динамических процессов и прагматики документов.

Таким образом, актуальность диссертационной работы определяется тем, что предлагаемая в работе модель:

а) относится к направлению Semantic Web (использует нотацию языка XML);

б) поддерживает представление динамических явлений и прагматики документов, что является новым для семантических моделей с нотацией языка XML;

Цель работы. Целью диссертационной работы является создание семантико-прагматической модели документа (СПМД) в нотации языка XML, расширяющей ряд существующих алгоритмов семантического анализа документов.

В процессе достижения данной цели были сформулированы и решены следующие задачи:

- формирование системы семантических понятий, обеспечивающей моделирование процессов и явлений и, тем самым, поддерживающей анализ динамики процессов;
- формирование системы прагматических понятий, поддерживающей анализ целей авторов документов и средств выражения этих целей;
- формализация СПМД на языке XML;
- разработка методики формирования смысловых представлений документов с использованием СПМД;
- разработка типовых алгоритмов семантического анализа смысловых представлений документов и проектирование языка запросов к XML-базам данных, хранящим смысловые представления.

Методы исследования. В работе использовались: теория множеств (для формализации понятий, лежащих в основе модели); семиотика (понятия документа, семантической и прагматической моделей, некоторые другие ключевые понятия работы основаны на представлении о знаковых системах); системный анализ, как базовая методологическая концепция, лежащая в основе исследования.

Научная новизна работы заключается в следующем:

- метод представления движения объекта в виде кортежа его состояний и переходов адаптирован к моделированию естественно-языкового описания движения объекта на уровне семантики и синтаксиса (расширены понятия перехода и состояния; сетевая нотация преобразована в иерархическую (XML));
- предложенная модель, включает в себя, помимо семантических элементов, и прагматические элементы, что позволяет описывать не только смысловое содержимое документа, но, так же, цели автора документа и ассоциированные с этими целями языковые средства (типы иллокутивных актов), которые он использовал;
- формирование смысловых представлений документов на базе СПМД осуществляется в процессе синтеза этих документов, что дает возможность использовать семантические шаблоны и обойти использование лингвистических процедур (в частности, синтаксического анализа), уровень развития которых недостаточен для формирования полноценных семантико-прагматических представлений документов.

Практическое значение. Метод формирования смысловых представлений документов, основанный на использовании предложенной в работе модели, позволяет параллельно процессу синтеза документов формировать XML-базу данных, которая является основой для реализации:

– алгоритмов семантического анализа, направленных на исследование динамики явлений (в частности, анализа динамики отношений между объектами; окружения объектов; анализа истории изменений объектов);
– поисковых запросов, осуществляющих семантический поиск не на уровне синтаксических конструкций, а на уровне глубинного смысла.

В процессе написания диссертационной работы, создана и внедрена в нескольких организациях система управления документами, поддерживающая метод построения баз данных смысловых представлений документов на основе СПМД. Все элементы данного метода продемонстрировали свою работоспособность. Наибольшую эффективность своей работы система продемонстрировала в организациях, для которых характерен интенсивный процесс создания документов (в частности, в нотариальных конторах).

Апробация работы. Основные положения диссертации и результаты исследований излагались на конференциях и семинарах, в том числе в рамках Международной конференции “Информационные технологии в управлении промышленностью и экономикой субъектов РФ” (г. Челябинск, 2002); Четвертой Всероссийской научной internet-конференции “Компьютерные технологии и моделирование в естественных науках и гуманитарной сфере”; межотраслевой научно-практической конференции “Снежинск и наука” (г. Снежинск); межвузовской научно-практической конференции "Автоматизация и информационные технологии” (г. Набережные Челны).

Связь с государственными программами. Диссертационная работа связана с тематикой работ, осуществляемых в соответствии с Федеральной целевой программой “Электронная Россия”.

Публикации. Базовые положения диссертации отражены в 8 публикациях.

Структура и объем работы. Диссертационная работа состоит из введения, пяти глав, заключения, списка литературы и трех приложений. Общий объем работы составляет 207 страниц (в том числе приложения – 51 страница). В работу входит 38 рисунков, 7 таблиц. Список литературы содержит 65 наименований.

На защиту выносятся следующие основные положения:

1) семантико-прагматическая модель в нотации XML, поддерживающая представления динамических явлений и прагматики документов;

2) методика автоматического формирования смысловых представлений документов на базе предложенной модели в процессе их синтеза с использованием синтаксических и семантических шаблонов;

3) типовые алгоритмы семантического анализа смысловых представлений документов: анализ динамики отношений, окружения и истории изменений объектов;

4) язык запросов к XML-базам данных, хранящим смысловые представления документов, построенные на основе предложенной семантико-прагматической модели документа.

ОСНОВНОЕ СОДЕРЖАНИЕ РАБОТЫ

Ключевые понятия диссертационной работы

Документ – информация, зафиксированная на материальном носителе в виде семиотической (знаковой) системы и имеющая выраженную прагматическую (целевую) направленность в некоторой предметной области;

Как информация, зафиксированная с помощью знаковой системы, документ может быть формально представлен в соответствии с определением семиотической системы (Поспелов Д.А.):

$$C = <M, x_T, x_P, x_A, x_П, x_Z, x_D, x_V>, \quad (1)$$

где M – формальная интерпретированная модель, то есть модель вида

$$M = \{T, P, A, П\}, \quad (2)$$

где T – множество базовых элементов (например, слов естественного языка); P – множество синтаксических правил (правил комбинации базовых элементов); A – множество аксиом (допустимых синтаксических конструкций, имеющих некоторый смысл, т.е. ассоциированных с некоторым множеством концептов, денотатов и целей); П – семантические правила (правила построения из аксиом – новых синтаксически допустимых и осмысленных конструкций), для которой задана модель интерпретации

$$L = \{Z, D, V\} \mid D \supset A, V \supset П, \quad (3)$$

где Z – множество интерпретируемых значений (множество концептов, денотатов и целей, которые могут быть ассоциированы с допустимыми синтаксическими конструкциями); D – множество правил отображения T -> Z (правил ассоциации базовых элементов с концептами, денотатами, целями); V – правила интерпретации, позволяющие приписывать семантической совокупности, состоящей из уже интерпретированных элементов, некоторое значение из Z (правила ассоциации семантических конструкций с концептами, денотатами и целями);

$x_T, x_P, x_A, x_П$ – соответственно правила изменения T, P, A, П;

x_Z, x_D, x_V – соответственно правила изменения Z, D, V.

Под **интерпретацией знаковой системы** (формальной модели M) понимается ассоциация знаков с концептами, денотатами и целями (отображение множества T на множество Z), в соответствии с правилами интерпретации (элементами множеств A, П, D, V).

Семантическая модель (модель семантики) – формализованное представление модели интерпретации L, устраняющее гибкость и неоднозначность последней, и, тем самым, позволяющее автоматизировать семантический анализ знаковых систем:

$$L' = \{Z', D', V'\} \mid \nexists x_{Z'}, \nexists x_{D'}, \nexists x_{V'} \quad (4)$$

Формализация L достигается с одной стороны за счет устранения гибкости модели интерпретации (исключения x_Z, x_D, x_V), а с другой стороны –

упрощения множества концептов, денотатов и целей – Z. Упрощение принято осуществлять путем выделения в множестве Z базовых классов и сведения его к ним (Z -> Z′ | Z′⊂ Z). С семантической моделью L′ ассоциируется некоторая знаковая система M′, в результате чего получается семиотическая система C′ (язык моделирования, язык представления знаний), с помощью которой можно представлять смысл документов и автоматизировать семантический анализ.

Решение задач диссертационной работы

Цель, поставленная в данной диссертационной работе (разработка метода формирования смысловых представлений документов на основе СПМД), определяет три ключевых задачи:

- синтез в нотации XML семантико-прагматической модели, которая по своему элементному составу и структуре обеспечивала бы представление динамических явлений и прагматики документов;
- разработка метода формирования смысловых представлений реальных документов на основе данной модели;
- разработка типовых алгоритмов семантического анализа модели и языка запросов к базам данным, хранящим смысловые представления.

Первым шагом решения задачи синтеза модели был анализ уже существующих семантических и прагматических моделей: “Компонентный анализ” Катца и Фодора, “Семантические падежи” Филмора, модель Уилкса – Мастермана, “Концептуальная зависимость” Шенка, “Смысл – текст” Жолковского, “Теория речевых актов” Остина, а так же некоторые языки моделирования и представления знаний и семантические модели исследовательского направления Semantic Web (в частности, RDF). На основе обобщения этих моделей был сделан следующий вывод: человек воспринимает и представляет окружающий мир в виде объектов, характеризуемых свойствами и отношениями между собой или, другими словами, – своими состояниями. Состояния объектов находятся в непрерывном изменении. Изменения происходят во времени и в соответствии с причинно-следственными закономерностями. Таким образом, модель, поддерживающая представление динамики, должна представлять собой описание временной и причинно-следственной составляющих изменения свойств и отношений взаимодействующих между собой объектов. В связи с этим было сформировано множество понятий, лежащих в основе элементного состава и структуры синтезируемой модели (понятия описаны формально с помощью аппарата теории множеств):

- свойство – кортеж вида:

$$P_n = \langle n, v_{cur} \rangle \mid v_{cur} \in \{v_1, v_2, \ldots, v_k, \ldots\}, \quad (5)$$

где n – имя свойства,
v_{cur} – текущее значение свойства;
$\{v_1, v_2, \ldots, v_k, \ldots\}$ – множество допустимых значений свойства.

- объект – множество свойств, такое что

  а) в универсальном множестве объектов нет в данный момент времени другого объекта, абсолютно идентичному данному;

  б) данный объект связан как минимум с одним отличным от него объектом;

  в) множество свойств и/или их значений изменяются с течением времени (объект движется).

  Формально говоря:

$$O_i^{tn} = \{P_1, P_2, \ldots, P_k, \ldots\} \mid \quad (6)$$

$$\forall (O_i^{tn}, O_j^{tn}) \mid O_i^{tn} \in U^O, O_j^{tn} \in U^O \rightarrow$$
$$\exists (P_l^{Oi}, P_m^{Oj}) \mid (P_l^{Oi}, P_m^{Oj}) \in O_i^{tn} \times O^{tn}_j,\ l \neq m \vee P_l^{Oi} \neq P_m^{Oj};$$
$$\exists O_j^{tn} \mid O_j^{tn} \in U^O, O_j^{tn} \neq O_i^{tn} \rightarrow (O_i^{tn}, O_j^{tn}) \in R^{tn}, R^{tn} \subseteq U^O \times U^O;$$
$$\forall (O_i^{tm}, t_m) \mid (O_i^{tm}, t_m) \in O_i \times T \rightarrow$$
$$\exists (O_i^{tn}, t_n) \mid (O_i^{tn}, t_n) \in O_i \times T,\ n > m,\ O_i^{tm} \neq O_i^{tn},$$

  где U^O – универсальное множество объектов в момент времени t_n;
  $T = <\ldots, t_{n-1}, t_n, t_{n+1}, \ldots>$ – время;
  $O_i = \{O_i^{t1}, O_i^{t2}, \ldots, O_i^{tn}, \ldots\}$ – объект на протяжении всего времени его существования (множество состояний объекта).

- отношение – изменяемое во времени множество пар объектов (участвующих в отношении или связанных отношением):

$$R^{tn} = \{(O_i^{tn}, O_j^{tn}) : (O_i^{tn}, O_j^{tn}) \in U^O \times U^O\} \mid \quad (7)$$

$$\forall (R^{tm}, t_m) \mid (R^{tm}, t_m) \in R \times T \rightarrow$$
$$\exists (R^{tn}, t_n) \mid (R^{tn}, t_n) \in R \times T,\ n > m,\ R^{tm} \neq R^{tn},$$

  где U^O – универсальное множество объектов;
  T – время;
  $R_i = \{R^{t1}, R^{t2}, \ldots, R^{tn}, \ldots\}$ – отношение на протяжении всего времени его существования.

- взаимодействие объектов – взаимное изменение во времени свойств объектов и отношений, в которых они состоят:

$$(O_i^{tm} \rightarrow O_i^{tn}), (R^{tm} \rightarrow R^{tn}) \rightarrow (O_j^{tn} \rightarrow O_j^{tp}), (R^{tn} \rightarrow R^{tp}) \rightarrow \quad (8)$$
$$\rightarrow (O_i^{tp} \rightarrow O_i^{ts}), (R^{tp} \rightarrow R^{ts}),$$

  где $(O_i^{tn} \rightarrow O_i^{tm})$ – переход объекта из одного состояния в другое;
  $(R^{tm} \rightarrow R^{tn})$ – изменение отношения R.

- причина, следствие – взаимные изменения свойств и отношений объектов, одни из которых (изменения-причины) обязательно влекут за собой другие (изменения-следствия)

$$\forall\ t_m \mid (O_i^{tm} \to O_i^{tn}), (R^{tm} \to R^{tn}) \to\ (O_j^{tn} \to O_j^{tp}), (R^{tn} \to R^{tp}), \qquad (9)$$

где n – m = const, p – n = const; $(O_i^{tm} \to O_i^{tn}), (R^{tm} \to R^{tn})$ – причина;
$(O_j^{tn} \to O_j^{tp}), (R^{tn} \to R^{tp})$ – следствие.

- система – составной объект, для которого выполняется два принципиальных условия: его свойства не сводимы к сумме свойств составляющих его объектов; некоторый набор его свойств из всего множества свойств – постоянен (объект относительно стабилен)

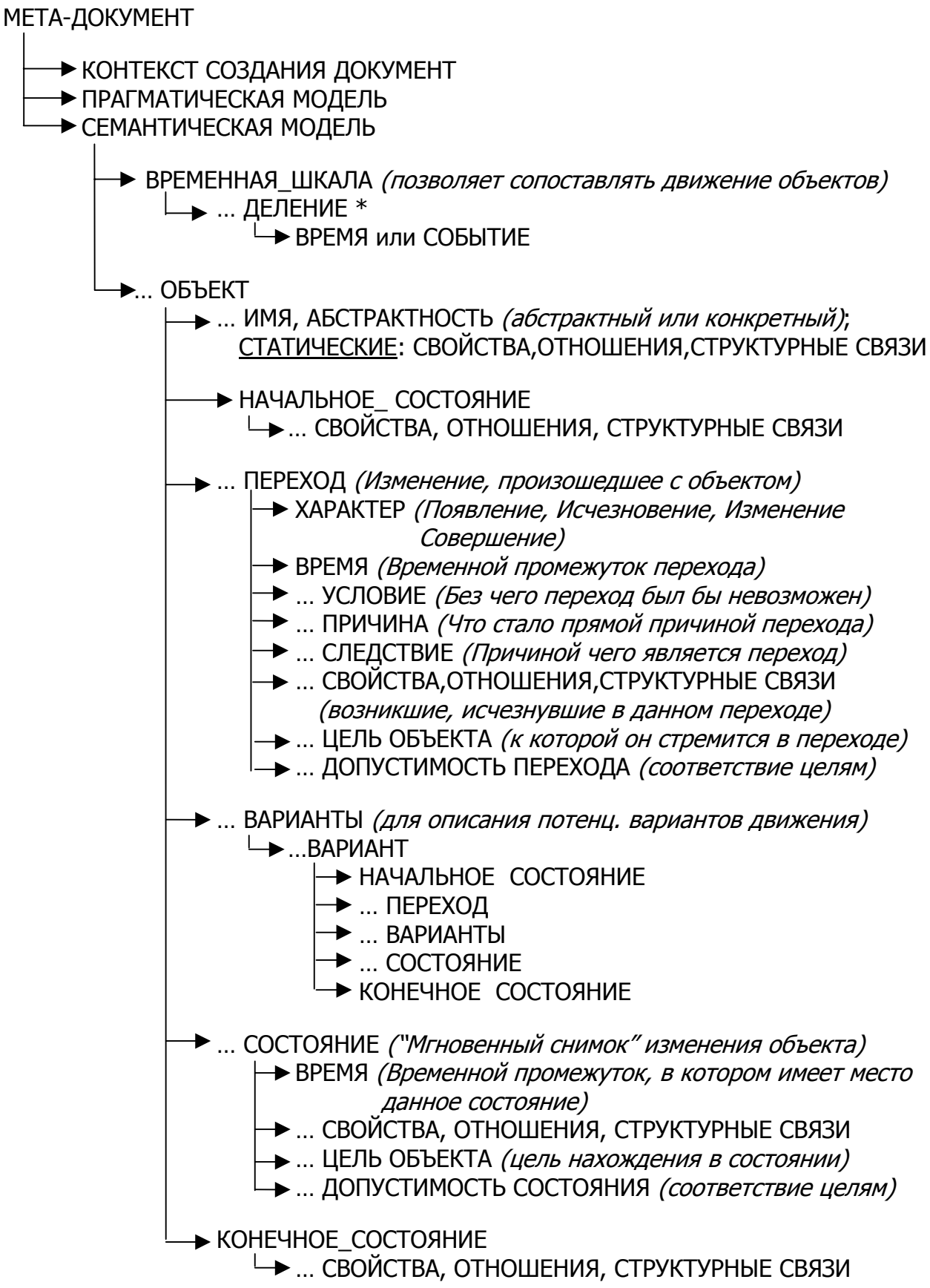
$$\begin{cases} S = \{E_1, E_2, \ldots, E_i, \ldots, R_s\},\ E_i - \text{объект (элемент)} \\ R_s \subseteq \{E_1, E_2, \ldots, E_i, \ldots\}^2 \\ S \neq \bigcup\limits_i E_i \\ \forall\ (t_n, t_m) \mid t_n \in T,\ t_m \in T,\ n > m \to \exists\ S_{const} \mid S_{const} \subseteq S,\ S^{tn}{}_{const} = S^{tm}{}_{const.} \end{cases} \qquad (10)$$

Данная система понятий легла в основу элементного состава синтезируемой модели. Следующим шагом синтеза модели стало формирование структуры, обеспечивающей представление динамических явлений.

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 мат. 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. Полученная структура показана на рис.1.

Как видно из этого рисунка, элементами верхнего уровня семантической модели являются элементы “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 (с помощью элементов “деление”). 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.



* "→…" означает, что элемент может присутствовать 0, 1 и более раз

Рис.1. Структура синтезируемой модели

Объект характеризуются 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 “объект1 – объект2”, 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 моделируемого явления.

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 (переходы) 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.

Модель описывает последовательность переходов и состояний объекта, которые имели место в действительности. Однако может потребоваться описать и потенциальные варианты движения объекта, которые могли быть (могут быть в настоящем или будущем). Более того, такие потенциальные состояния и переходы могут быть условиями и причинами выбора объектом реального варианта движения. Для описания потенциальных переходов и состояний в модель введен элемент “Варианты”.

За формированием семантических элементов и динамической структуры модели последовало формирование системы прагматических элементов, которые позволяют в дополнение к семантическим понятиям, описывающим содержание документа, описывать характеристики авторов и адресатов документов, цели авторов в отношении адресатов и лингвистические средства, которыми авторы пользовались для достижения этих целей. Ключевыми понятиями множества прагматических элементов модели являются два понятия:

- участники общения – авторы и адресаты документа, характеризующиеся своими целями, социальным статусом, знаниями, эмоциональным состоянием, уверенностью;
- прагматический блок: автор передает адресату информацию с определенной целью; в достижении этой цели участвует не только сама передаваемая информация, но и способ ее языковой организации (тип иллокутивного акта); так, автор одну и ту же информацию может передать в форме утверждения, предположения, вопроса, приказа и т.д.; отсюда возникает понятие прагматического блока – части документа, отличной от других подобных частей определенным способом языковой организации – типом иллокутивного акта; множество прагматических блоков составляют прагматическую модель документа:

$$D = <(B_1, k_1, p_1), (B_2, k_2, p_2), \ldots, (B_i, k_i, p_i), \ldots>, \qquad (11)$$

где B_i – прагматический блок (некоторая часть документа);

$k_i \in$ {Побуждение к действию, Разрешение, Запрещение, Ограничение, Утверждение, Вопрос, Предположение, Одобрение, Неодобрение, Обещание, Обязательство, Оспаривание, Протест, Сожаление, Благодарность, Объяснение, Изложение} – иллокутивный тип *i*-го прагматического блока;

$p_i \in$ {Изменение физического состония, Изменение эмоционального состония, Изменение мнения, Передача информации} – цель автора в *i*-м прагматическом блоке.

Таким образом, предлагаемая модель, как показано на рис.1., состоит из трех подмоделей – семантической, прагматической и подмодели описания контекста создания документа (в частности описания авторов и адресатов). Она формализована в виде XML-схем.

Рассмотрим пример моделирования небольшого сообщения: *“Появились слухи. что после того, как господин X возглавил корпорацию A, он уволился из корпорации B, поскольку не мог совмещать одновременно две руководящих должности. Доподлинно известно, что он переехал из города C в город D”*.

Графическая нотация семантической модели представлена на рис.2. (Здесь кругами обозначаются состояния; прямоугольниками – переходы [/–возникновение; \ – прекращение; \/\ – изменение]; линиями с точками на концах – отношения; двойными линиями – причинно-следственные связи).

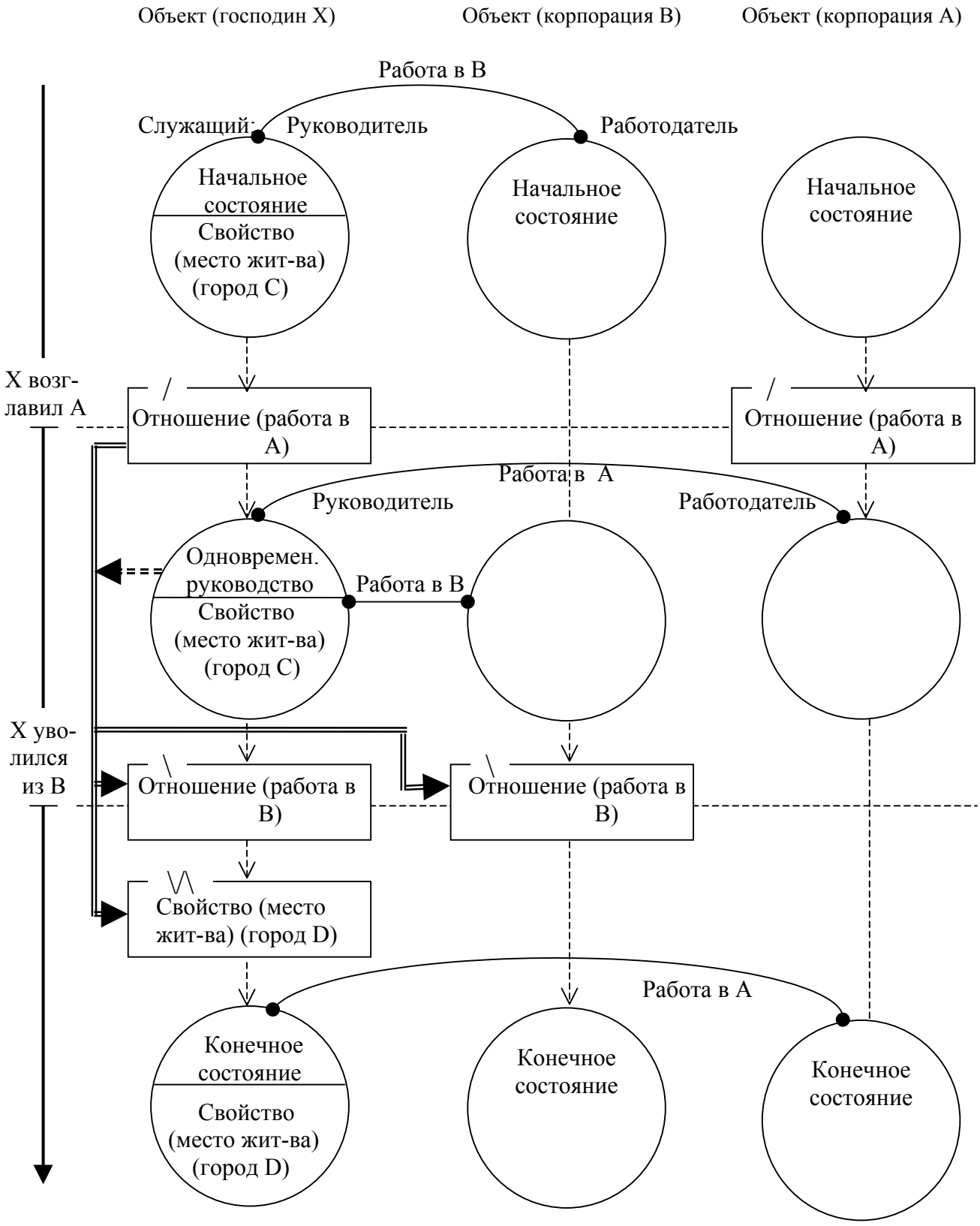


Рис.2. Графическая нотация семантической модели

Прагматическая модель проиллюстрирована табл. 1.

Таблица 1

№	Название прагма-блока	Текст	Цель	Средство
1	Появились слухи	Появились слухи. что после того, как господин X возглавил корпорацию A, он уволился из корпорации B, поскольку не мог совмещать одновременно две руководящих должности.	– Передача информации – Изменение эмоционального состояния	Предположение
2	Доподлинно известно	Доподлинно известно, что он переехал из города C в город D	– Передача информации – Изменение эмоционального состояния	Утверждение

СПМД документа в нотации XML (в сокращенном виде – здесь описано движения только одного объекта из трех):

```
<?xml version="1.0" encoding="cp866"?>
<МЕТА-ДОКУМЕНТ>

<ПРАГМАТИЧЕСКАЯ_МОДЕЛЬ>
    <ПРАГМА-БЛОК Имя=" Появились слухи">
        <ЦЕЛЬ Тип="Передача_информации"/>
        <ЦЕЛЬ Тип="Изменение_эмоц_состония"/>
        <СРЕДСТВО Тип="Предположение"/>
        <ТЕКСТ> Появились слухи. что после того, как X возглавил корпорацию A,
                он уволился из корпорации B, поскольку не мог совмещать
                одновременно две руководящих должности.
        </ТЕКСТ>
     </ПРАГМА-БЛОК>

    <ПРАГМА-БЛОК Имя="Доподлинно известно">
        <ЦЕЛЬ Тип="Передача_информации"/>
        <ЦЕЛЬ Тип="Изменение_эмоц_состония"/>
        <СРЕДСТВО Тип="Утверждение"/>
        <ТЕКСТ> Доподлинно известно, что он переехал из города C в город D
        </ТЕКСТ>
     </ПРАГМА-БЛОК>
</ПРАГМАТИЧЕСКАЯ_МОДЕЛЬ>

<СЕМАНТИЧЕСКАЯ_МОДЕЛЬ>
  <ВРЕМЕННАЯ_ШКАЛА>
        <ДЕЛЕНИЕ> <СОБЫТИЕ Значение="X возглавил A"/>
        </ДЕЛЕНИЕ>
        <ДЕЛЕНИЕ> <СОБЫТИЕ Значение="X уволился из B"/>
        </ДЕЛЕНИЕ>
  </ВРЕМЕННАЯ_ШКАЛА>

  <ОБЪЕКТ Имя="господин X">
     <НАЧАЛЬНОЕ_СОСТОЯНИЕ>
          <СВОЙСТВО UID="Место_жительства_до_переезда">
```

```
        <ИМЯ>Место жительства</ИМЯ>
        <ЗНАЧЕНИЕ>город C</ЗНАЧЕНИЕ>
    </СВОЙСТВО>
    <ОТНОШЕНИЕ Имя=”Работа в B” UID=”Работа в B”>
      <РОЛЬ>Служащий</РОЛЬ> <РОЛЬ>Руководитель</РОЛЬ>
      <УЧАСТНИК Объект=”корпорация B”/>
    </ОТНОШЕНИЕ>
</НАЧАЛЬНОЕ_СОСТОЯНИЕ>

<ПЕРЕХОД Характер=”Появление” UID=”X начал работу в A”>
     <ВРЕМЯ>
       <В_МОМЕНТ Измерение=”Событие” Значение=”X возглавил A”/>
     </ВРЕМЯ>
     <ОТНОШЕНИЕ Имя =”Работа в A” UID=”Работа в A”>
        <РОЛЬ>Руководитель</РОЛЬ> <УЧАСТНИК Объект=” корпорация A”/>
     </ОТНОШЕНИЕ>
</ПЕРЕХОД>

<СОСТОЯНИЕ UID=”Одновременное руководство”>
    <ДОПУСТИМОСТЬ Значение=”Отрицательно”
                          UID=”Запрет на совмещение”/>
    <СВОЙСТВО Ссылка=”Место_жительства_до_переезда”/>
    <ОТНОШЕНИЕ Имя=”Работа в A” Ссылка=”Работа в A”/>
    <ОТНОШЕНИЕ Имя=”Работа в B” Ссылка=”Работа в B”/>
</СОСТОЯНИЕ>

<ПЕРЕХОД Характер=”Прекращение” UID=”X уволился из B”>
     <ВРЕМЯ>
       <В_МОМЕНТ Измерение=”Событие” Значение=”X уволился из B”/>
     </ВРЕМЯ>
     <УСЛОВИЕ>
        <ССЫЛКА UID=”Одновременное руководство”/>
     </УСЛОВИЕ>
     <ПРИЧИНА>
        <ССЫЛКА UID=”Работа в A”/>
     </ПРИЧИНА>
    <ОТНОШЕНИЕ Имя=”Работа в B” UID=”Увольнение из B”>
      <РОЛЬ>Служащий</РОЛЬ> <РОЛЬ>Руководитель</РОЛЬ>
      <УЧАСТНИК Объект=”корпорация B”/>
    </ОТНОШЕНИЕ>
</ПЕРЕХОД>

<ПЕРЕХОД Характер=”Изменение” UID=”X сменил место жительства”>
     <ВРЕМЯ>
       <ПОСЛЕ Измерение=”Событие” Значение=”X уволился из B”/>
     </ВРЕМЯ>
     <ПРИЧИНА>
       <ССЫЛКА UID=”Работа в A”/>
     </ПРИЧИНА>
    <СВОЙСТВО UID=”Место_жительства_после_переезда”>
       <ИМЯ>Место жительства</ИМЯ>
       <ЗНАЧЕНИЕ>город D</ЗНАЧЕНИЕ>
    </СВОЙСТВО>
</ПЕРЕХОД>

<КОНЕЧНОЕ_СОСТОЯНИЕ>
  <СВОЙСТВО Ссылка=”Место_жительства_после_переезда”/>
```

```
      <ОТНОШЕНИЕ Имя="Работа в А" Ссылка="Работа в А"/>
    </КОНЕЧНОЕ_СОСТОЯНИЕ>
  </ОБЪЕКТ>
</СЕМАНТИЧЕСКАЯ_МОДЕЛЬ>
</МЕТА-ДОКУМЕНТ>
```

Условие применимости семантико-прагматической модели документа в некоторой предметной области может быть сформулировано на основе определения документа, как семиотической системы (1):

Пусть дано непустое множество документов, составляющих данную предметную область,

$$U^C = \{C_1, \ldots, C_i, \ldots\} \mid U^C \neq \varnothing,$$

зададим отношение R на множестве $U^C \times U^C$:

$$(C_i, C_j) \in R, \text{ если } T_i = \{T_i^1, T_i^2\} \mid |T_i^1| > |T_i^2|; \quad (12)$$
$$T_j = \{T_j^1, T_j^2\} \mid |T_j^1| > |T_j^2|;$$
$$T_i^1 = T_j^1 \text{ и } P_i = P_j,$$

где, T_i^1, T_j^1 – множества регулярных (не изменяющихся от одного документа подмножества к другому) базовых элементов *i*-го и *j*-го документов;
T_i^2, T_j^2 – множества нерегулярных (изменяющихся от одного документа данного подмножества к другому) базовых элементов *i*-го и *j*-го документов.

Если данное отношение R является отношением эквивалентности на множестве $U^C \times U^C$ (разбивает его на такие подмножества, что каждый документ из множества U^C относится к одному и только одному подмножеству), т.е., если документы данной предметной области могут быть разбиты на множества, элементы каждого из которых удовлетворяют единому синтаксическому шаблону, то СПМД применима к данной предметной области U^C.

Центральной проблемой, связанной с применением предложенной СПМД является проблема автоматического формирования смысловых представлений конкретных документов на ее основе. Формирование смысловых представлений на основе уже готового документа не представляется на данный момент возможным, поскольку природа различия поверхностных и глубинных семантических моделей не ясна. Таким образом, единственным источником модели остается эксперт. Тем не менее, определенный уровень автоматизации – возможен. В частности, можно использовать смысловой шаблон, построенный экспертом для ряда однотипных документов, и синтезатор текста, работающий на основе синтаксического шаблона этих документов. Для реализации данной идеи в язык описания синтаксических шаблонов документов необходимо включить специальные элементы, отвечающие за связь полей синтаксического и смыслового шаблонов. Тогда оператор, создающий новый документ в синтезаторе текста, заполняя поля синтаксического шаблона, будет параллельно формировать смысловое представление. Полный цикл применения СПМД показан на рис.3.

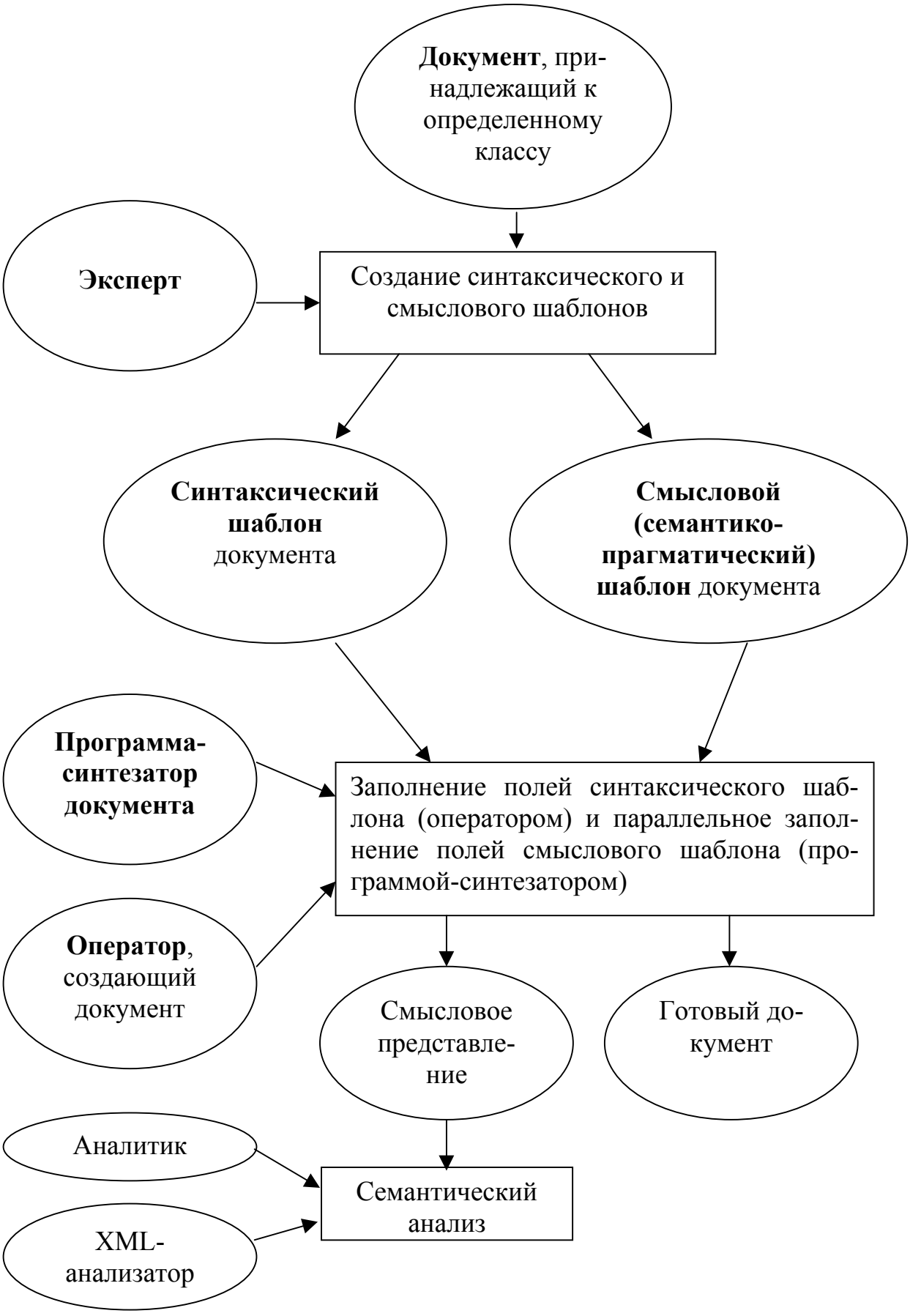


Рис.3. Схема работы с моделью

Язык описания синтаксических шаблонов, используемый для формирования моделей документов, состоит из общих команд, используемых для автоматизации синтеза документа, и команд, отвечающих за формирование смыслового представления. РБНФ - описание языка приведено в диссертационной работе.

Общие команды делятся на команды описания полей и команду выбора. Поле представляет переменный текст – текст, который меняется от документа к документу (ФИО, даты и т.п.). Программа - синтезатор текста, встретив в бланке поле, генерирует на экране специальное поле ввода, в котором оператор может набрать соответствующий переменный текст (либо программа-синтезатор сама вставляет в поле ввода некоторый текст, в зависимости от команды поля). Команды поля служат для автоматизации его заполнения. Они делятся на три группы: команды вставки даты, команды вставки текста и команды склонения фамилий. Команда выбора обеспечивает гибкость языка шаблонов, необходимую для описания документов, содержимое которых меняются от ситуации к ситуации достаточно сильно. При обработке данной команды, синтезатор текста выдает на экран окно выбора, в котором оператор выбирает один из возможных вариантов шаблона.

Команда “Мета-документ”, отвечающая за формирование семантической модели, используется, как контейнер семантико-прагматического (смыслового) шаблона, построенного на основе СПМД. Там, где в смысловом шаблоне должна присутствовать конкретная информация (фамилии, адреса и т.п.), стоят ссылки на соответствующие поля синтаксического шаблона – команды вставки поля. В момент заполнения шаблона команда “Мета-документ” игнорируется синтезатором текста. Работать эта команда начинает лишь в момент сохранения оператором уже набранного документа. При этом на место команд вставки поля подставляются значения соответствующих полей. Далее модель сохраняется в отдельном файле, имя которого совпадает с именем созданного документа, а расширение – “xml”.

Третья ключевая задача диссертационной работы – разработка типовых алгоритмов семантического анализа модели и языка запросов к ней.

В рамках данного исследования разработаны несколько типовых алгоритмов семантического анализа (все они реализованы на языке VBA и успешно протестированы).

– **Анализ отношений объекта**. На практике, данный 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)
– **Анализ окружения объекта**. Этот анализ дает возможность узнать – с кем общался (состоял в отношениях, в отношении кого совершал действия) интересующий объект. Пример отчета по результатам данного анализа приведен на рис.3.

– **Анализ истории изменений объекта**. Этот анализ дает возможность проследить ретроспективу реальных изменений, произошедших с интересующим объектом за определенный период.

Двумя основными аспектами **языка запросов к СПМД**, являются:

– множество сущностей (элементов модели), которые можно извлекать из модели;
– система шаблонов (wildcards), позволяющих строго и нестрого описывать искомые элементы модели.

Эти аспекты описаны в схеме языка, формализованной при помощи нотации РБНФ. Схему языка можно увидеть в диссертационной работе.

Далее приведен пример извлечения информации из XML-базы данных, хранящей смысловые представления, с одновременным применением анализа истории изменений объекта. Пусть требуется извлечь информацию об истории купли-продажи квартиры, находящейся по адресу: город A, улица B, дом C, № D.

Поисковый запрос:

ВЫБРАТЬ ОБЪЕКТ.ИСТОРИЯ (ИМЯ='Квартира' И СВОЙСТВО (ИМЯ='Адрес', ЗНАЧЕНИЕ='*A*B*C-D'))

Результат выполнения запроса показан в табл. 2.

Таблица 2

ИСТОРИЯ ИЗМЕНЕНИЙ ОБЪЕКТА "Квартира":
Всего просмотрено моделей: 50; Из них в анализ включено: 2
Дата самого раннего документа: 20.09.2001; Дата самого позднего документа: 23.09.2001

Дата	Период	Изменение	Участники изменения	Причина изменения
20.09.2001	В_МОМЕНТ (заключение договора)	Объект (Объект_действия) подвергся действию "Купля-Продажа"	Участник X; Участник Y; Нотариус	
	ПОСЛЕ (заключение договора); ДО (регистрация договора)	Объект (Объект_действия) подвергся действию "Передача квартиры"	Участник X; Участник Y	Совершение действия "Купля-Продажа"
	ПОСЛЕ (передача квартиры); ДО (регистрация договора)	Объект (Объект собственности) изменил характер отношения "Владение квартирой"	Участник Y	Совершение действия "Передача квартиры"
23.09.2001	В_МОМЕНТ (заключение договора)	Объект (Объект_действия) подвергся действию "Купля-Продажа"	Участник Y; Участник Z; Нотариус	
	ПОСЛЕ (заключение договора); ДО (регистрация договора)	Объект (Объект_действия) подвергся действию "Передача квартиры"	Участник Y; Участник Z	Совершение действия "Купля-Продажа"
	ПОСЛЕ (передача квартиры); ДО (регистрация договора)	Объект (Объект собственности) изменил характер отношения "Владение квартирой"	Участник Z	Совершение действия "Передача квартиры"

ОСНОВНЫЕ ВЫВОДЫ И РЕЗУЛЬТАТЫ

Проведенные в диссертационной работе исследования образуют теоретическую и практическую основу для расширения ряда существующих алгоритмов семантического анализа документов и позволяют сформулировать следующие выводы и получить практические результаты:

1) анализ существующих семантических моделей и методик семантического анализа естественно-языковых текстов позволил сделать следующий вывод: на данный момент существует два основных подхода к решению задачи семантического анализа. Первый подход (лингвистический) ориентирован на построение смыслового представления текста путем проведения ряда лингвистических и статистических анализов: морфологического, синтаксического, частотного и др. Смысловое представление, формируемое в этом случае, представляет собой семантическую сеть (или ее разновидность), которая дает возможность реализовывать, главным образом, алгоритмы статического анализа (анализ наличия связей между объектами, силы этих связей и т.п.);

2) в рамках второго подхода анализируется не сам документ, а его смысловое представление, которое формулируется в искусственной синтаксической нотации на основе формальной семантической модели (например, семантическая модель web-документа RDF, использующая нотацию языка XML). Поскольку гибкий и сложный синтаксис естественного языка заменяется искусственной нотацией, а семантика естественного языка – формализуется, данный подход потенциально позволяет реализовывать значительно более сложные алгоритмы семантического анализа. Однако на данный момент в качестве формальных моделей используются те же семантические сети, что искусственно занижает возможности данного подхода. Необходимо создать новую формальную модель семантики документа;

3) синтезирована семантико-прагматическая модель документа в нотации языка XML. Данная модель:

- поддерживает представление динамических явлений;
- поддерживает представление целей авторов и средств их выражения;
- лежит в основе реализации новых алгоритмов семантического анализа, связанных с анализом динамики событий и прагматики документов;

4) разработана методика автоматического формирования смысловых представлений документов в процессе их синтеза, основанная на использовании шаблонов, созданных экспертами для различных классов документов на базе СПМД. Данный метод позволяет обойтись без использования лингвистических процедур анализа документов;

5) разработаны алгоритмы, которые расширили ряд существующих алгоритмов семантического анализа на базе использования СПМД (в частности, алгоритмы анализа динамики отношений между определенной группой объектов; анализа истории изменений, происходящих с интересующим объектом; анализа окружения объекта). Спроектирован язык запросов, позво-

ляющий применять вышеуказанные алгоритмы для анализа XML-баз данных, хранящих смысловые представления документов;

6) в процессе написания диссертационной работы, создана и внедрена в нескольких организациях система управления документами, поддерживающая предложенную в работе методику. Все элементы методики продемонстрировали свою работоспособность. Наибольшую эффективность своей работы система продемонстрировала в организациях, для которых характерен интенсивный процесс создания документов (в частности, в нотариальных конторах).

Публикации по теме диссертационной работы:

1. Коровин С.Е. Семантико-прагматическая модель документа в нотации XML// Электронный журнал "Исследовано в России" – 123 – С. 1360–1380, 2002 г. http://zhurnal.ape.relarn.ru/articles/2002/123.pdf
2. Кафтанников И.Л., Коровин С.Е. Семантика World Wide Web-документа // Вестник Юж.-Урал. гос. ун-та. Сер. Компьютер. технологии, упр., радиоэлектроника. – Вып. 1. – №9(09). – 2001. –С. 26–32.
3. Кафтанников И.Л., Коровин С.Е. Язык XML и семантика электронного документа: модель RDF//Интеллектика. Логистика. Системология: Сб. науч. тр.–Челябинск: Издание ЧНЦ РАЕН, РУО МАИ, ЧРО МАНПО, 2002.–Вып.7.–С.33–39.
4. Коровин С.Е. Язык XML и семантика электронного документа: динамическая семантическая модель //Интеллектика. Логистика. Системология: Сб. науч. тр.–Челябинск: Издание ЧНЦ РАЕН, РУО МАИ, ЧРО МАНПО, 2002.– Вып.7.–С.182–189.
5. Коровин С.Е., Мельников А.В., Кафтанников И.Л. Моделирование семантики и прагматики документа в нотации языка XML //Известия Челябинского научного центра. – Челябинск: Издание ЧНЦ УРО РАН, РФЯЦ – ВНИИТФ, ЮУрГУ, 2002, – Вып.16. –С.25–29.
6. Кафтанников И.Л., Коровин С.Е. Динамическая модель World Wide Web - документа //Сборник докладов межвузовской научно-практической конференции “Автоматизация и информационные технологии”, – Наб. Челны: Изд-во Камского госуд. политехн. ин-та, 2002.–С. 141–145.
7. Кафтанников И.Л., Коровин С.Е. Динамическая модель семантики WWW - документа //Труды Четвертой Всероссийской научной internet-конференции “Компьютерные технологии и моделирование в естественных науках и гуманитарной сфере”, – Тамбов: Изд-во ТГУ им. Г.Р.Державина, 2002. – Вып.22. –С. 35–39.
8. Коровин С.Е. Прагматическая модель WWW - документа //Труды четвертой Всероссийской научной internet-конференции “Компьютерные технологии и моделирование в естественных науках и гуманитарной сфере”, – Тамбов: Изд-во ТГУ им. Г.Р.Державина, 2002. – Вып.22. –С. 39–42.

Коровин Сергей Евгеньевич

ДИНАМИЧЕСКАЯ МОДЕЛЬ СЕМАНТИКИ И ПРАГМАТИКИ ДОКУМЕНТОВ НА БАЗЕ РАСШИРЕНИЯ ЯЗЫКА XML

Специальность 05.13.01 – "Системный анализ, управление и обработка информации (промышленность)"

Автореферат
диссертации на соискание ученой степени
кандидата технических наук

Издательство Южно-Уральского государственного университета

ИД №00200 от 28.09.99. Подписано в печать 17.10.2002. Формат 60x84 1/16.
Печать офсетная. Усл. печ. л. 1,16. Уч.-изд. л. 1. Тираж 80 экз. Заказ 267/415.

УОП Издательства. 454080, г. Челябинск, пр.им. В.И. Ленина, 76.