УДК 004.9: 343

М.К. Буза[1], Н.В. Деева[2]

ИНФОРМАЦИОННАЯ МОДЕЛЬ ПРОЦЕССА РАССЛЕДОВАНИЯ ПРАВОНАРУШЕНИЙ

Исследуется информационная модель следственного процесса в рамках обработки текстов протоколов допроса: анализируется протокол, выделяются его компоненты, проектируются алгоритмы их обработки. Предлагается алгоритм построения обобщенной объектной модели криминального дела, рассматриваются классы данных и средства их представления.

Введение

Современное состояние информационных технологий позволяет проектировать системы, которые упрощают работу человека в промышленной цепочке, где задачи хорошо формализованы, детерминированы и применяют четкую логику. Что касается задач, решаемых с помощью нетривиальных умозаключений (используя дедукцию, обобщение, интуицию и т. д.), то их построение с помощью стандартных средств и технологий не всегда представляется возможным. В такой области могут применяться экспертные системы или системы поддержки принятия решений. Они нацелены на выполнение рутинной вычислительной работы, анализируют данные и генерируют набор возможных альтернативных вариантов, а пользователь на их основе принимает решения.

Информационные технологии в последнее время находят широкое применение и в такой узкоспециализированной области, как криминалистика. В процессе изучения материалов дел следователь выполняет большую рутинную работу по оформлению различной документации, причем часть этой документации несет в себе информацию сугубо процессуального характера, регламентирующую лишь сам процесс расследования. Естественно, что следователь должен хранить в памяти большой объем данных, которые могут сыграть решающую роль в расследовании криминальных дел. Очевидным является то, что подобные виды деятельности можно и необходимо автоматизировать.

В связи с этим возникает задача разработки информационной модели следственного процесса с последующей ее реализацией средствами современных информационных технологий.

1. Типы уголовно-процессуальных документов

Анализ перечня документов [1], которые исследуются, обобщаются и готовятся в ходе следственного процесса и формируют уголовное дело, позволяет провести их *классификацию по данным* (рис. 1).

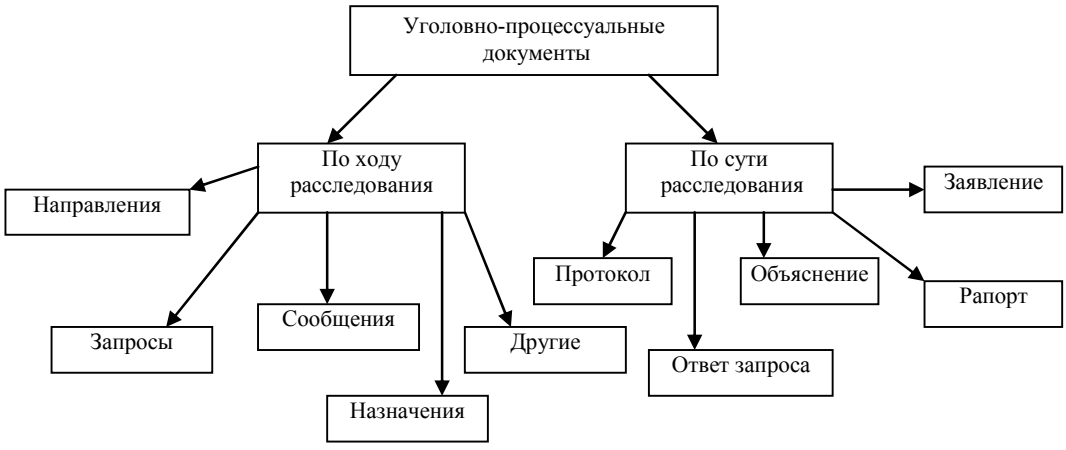

Рис.1. Типы уголовно-процессуальных документов

Отдельным классом документов, как правило, выделяются протоколы. *Протокол* – это акт, составленный уполномоченными на то должностными лицами, судебными или административными, для удостоверения тех или иных событий [2].

Выделим следующие типы протоколов: заявления, явки, допросы, осмотры, обыски, задержания, извлечения, освидетельствования, следственные эксперименты, очные ставки, предъявления, проверки показаний и др.

Объектом исследования данной статьи является набор уголовно-процессуальных документов, их структура, стилистические особенности и смысловой портрет текстов.

Предметом исследования является контент-анализ естественно-языковых текстов на морфосинтаксическом уровне: алгоритмы и методы извлечения смысла из текстов на естественном языке.

Одним из наиболее информативных является протокол допроса, который может характеризоваться по субъекту и характеру допроса. Параметры субъектов допросов и особенности допросов отражены в табл. 1.

Таблица 1

Характеристики протокола допроса

Протокол допроса											
Субъект допроса							Особенности допроса				
Возраст		Тип фигуранта дела					Характер допроса		Применение дополнительных устройств			
Совершеннолетний (по умолчанию)	Несовершеннолетний	Свидетель	Потерпевший	Обвиняемый	Подозреваемый	Эксперт	Первичный (по умолчанию)	Дополнительный	Без применения (по умолчанию)	Звукозапись	Видеозапись	Звуко- и видеозапись

Анализ классического протокола допроса позволяет выделить следующие основные части:

1. Заголовок – характеристика субъекта допроса по причастности к делу.
2. Информация о допросе – место, дата и время прохождения допроса.
3. Информация о следователе – данные о дознавателе, месте проведения допроса и статьях, на основании которых допрос был проведен.
4. Информация о субъекте допроса – краткая характеристика субъекта.
5. Статическая часть – сведения для субъекта допроса о формулировке дела, его правах и обязанностях.
6. Неизменяемая формулировка, предшествующая последующему рассказу субъекта.
7. Собственно рассказ субъекта допроса.
8. Вопрос следователя.
9. Ответ субъекта допроса.
10. Неизменяемая формулировка, как правило завершающая протокол допроса.
11. Подписи.

Выделим четыре типа компонентов протокола (рис. 2). Определим методы и алгоритмы обработки сегментов. Дадим определение каждому из сегментов и построим алгоритмы извлечения из них семантики.

Постоянные сегменты являются статическими юридическими формулировками, которые не влияют на общий ход расследования, но могут служить своеобразными маркерами для выделения других сегментов текста.

Процессуальные сегменты достаточно формализованы. Если предварительно сформировать шаблоны для следователя, то извлечение информации может проходить автоматически стандартными средствами.

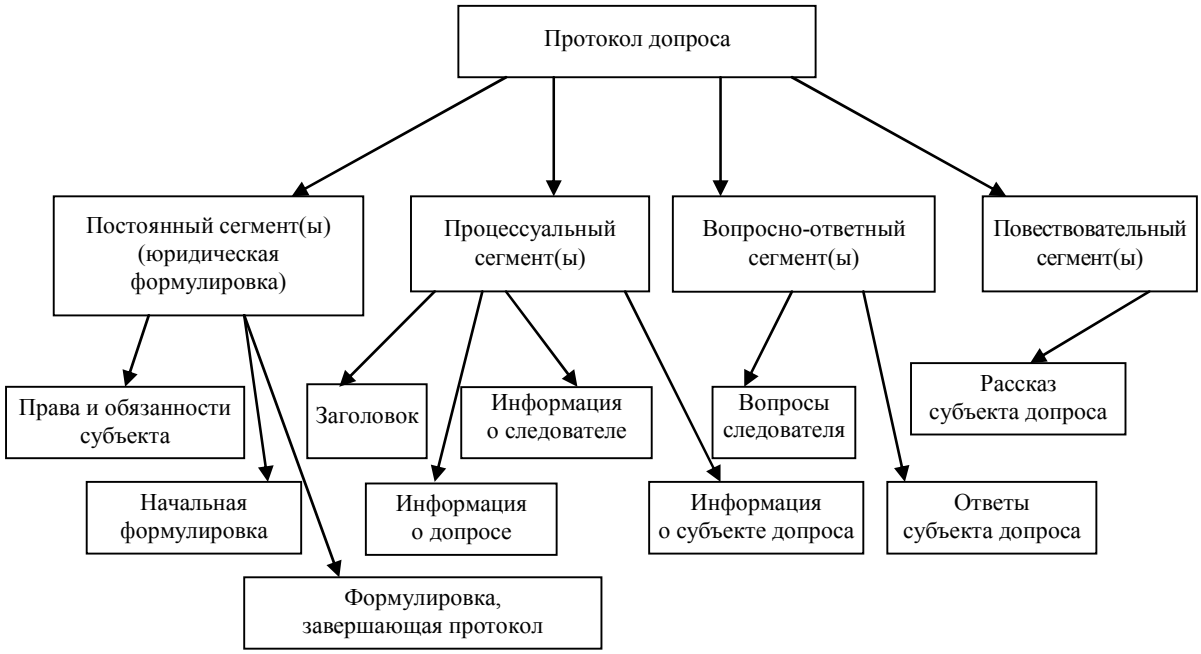

Рис. 2. Основные сегменты протокола допроса

Вопросно-ответные сегменты – это неформализованные сегменты, анализ которых требует специальных алгоритмов. Главной их особенностью является строгое следование ответа за соответствующим вопросом. Ключевые слова «Вопрос» и «Ответ» позволяют использовать их как маркеры, определяющие тип сегмента. Ответы характеризуются повествовательным стилем. Что касается вопросов, то их трактовка требует применения специальных алгоритмов.

Повествовательные сегменты – наиболее информативные части. Как правило, в документе встречается один такой сегмент. Он неформализованный, но смысл текста в нем в основном упорядочен по времени. Повествование происходит от одного лица и характеризуется отношением этого лица к действию повествования. Так как запись в протокол выполняет обычно следователь, то можно говорить, что стиль изложения материала не перегружен различного рода оборотами и смысловыми отступлениями.

2. Схема извлечения знаний из протоколов допроса

Наиболее интересной и трудоемкой задачей является анализ и извлечение смысла из повествовательных сегментов. Рассмотрим ее более подробно.

Для анализа и извлечения информации из повествовательного сегмента предложен следующий алгоритм:

1. Нормализация повествовательного сегмента, включающая:
– исключение анафорических ссылок (указательных и личных местоимений);
– обработку причастных оборотов;
– свертку сложных предложений;
– приведение имен собственных и числительных к общему заведомо определенному виду.
2. Синтаксический анализ сегмента (выделение предложений).

3. Морфологический анализ отдельного предложения (определение характеристик для каждого слова).

4. Синтаксический анализ предложения:

– выделение частей предложения (словосочетаний);

– построение новой формы предложения в виде $P = <p, s, a\_p, a\_s, d>$, где p – подлежащее, s – сказуемое, a_p – атрибуты подлежащего, a_s – атрибуты сказуемого, d – дополнение.

5. Построение объектной модели предложения:

– выделение действий;

– выделение субъектов действия;

– выделение объектов действия;

– выделение обстоятельств действия.

6. Построение объектной модели сегмента (выделение сущностей и их связей).

На основании предложенного алгоритма построим объектную «субъективную» смысловую картину некоторого фрагмента происшествия.

В целом для обработки текста протокола допроса необходимо выполнить:

1) анализ текста протокола на основе шаблона, который включает маркеры основных частей протокола и решает задачу сегментации протокола допроса;

2) извлечение знаний из процессуальных сегментов;

3) анализ и извлечение знаний из повествовательного сегмента;

4) анализ и извлечение знаний из вопросно-ответных сегментов;

5) построение общей объектной модели протокола допроса.

Определим последовательность и основные сущности, выделяемые на этапе анализа. Каждый этап обработки представим в виде программного модуля. Тогда анализ повествовательного сегмента будет включать следующую последовательность модулей:

1) создание и редактирование пакета документов;

2) сегментация текста протокола;

3) анализ повествовательного сегмента;

4) морфологический анализ предложения;

5) синтаксический анализ;

6) построение объектной модели предложения;

7) построение объектной модели текстового сегмента;

8) создание объектной модели протокола и обобщенной модели дела.

Представим указанные модули с определяемыми на данном этапе сущностями в виде схемы (рис. 3). Схема описывает этапы получения данных из текста и формирования базы знаний на основе композиции семантических сетей. Сущности, определяемые на каждом этапе, представляют собой объекты некоторых классов, которые, в свою очередь, могут создавать коллекции объектов. Под коллекцией будем понимать специализированные классы, использующие обычно единый интерфейс для хранения и извлечения данных.

В табл. 2 приведены типы классов, определяемые на каждом этапе; коллекции, которые объединяют объекты, а также алгоритмы и методы, используемые на каждом этапе.

Для реализации всех перечисленных в табл. 2 этапов разработаны программные модули. Рассмотрим функции каждого модуля.

Создание и редактирование пакета документов. На вход подаются текстовые документы, которые приобщаются к делу. Данный модуль представляет пользовательский интерфейс для пополнения, удаления и дальнейшей обработки материалов дела.

Сегментация текста протокола. На вход поступает текст протокола допроса. Модуль анализирует его и выдает размеченный текст с выделенными в нем сегментами. Для выделения смысловых частей в протоколе в качестве предопределенных маркеров используются постоянные (ключевые для протокола) сочетания: «сообщил», «вопрос», «ответ», «прочитан» и др.

Предлагается следующий алгоритм выделения сегментов:

1. Последовательно просматривая текст протокола допроса на базе шаблона и маркеров, выделяем в нем сегмент заголовка, сегмент информации о допросе, сегмент информации о следователе, сегмент информации о субъекте допроса, сегмент статической части.

2. Находим маркер «сообщил следующее».
3. Находим завершающий маркер – «сообщить нечего».
4. Выделяем текстовый сегмент.
5. Находим маркер «вопрос».
6. Находим маркер «ответ».
7. Выделяем вопросно-ответный сегмент.
8. Повторяем пп. 5–7, если не все вопросно-ответные сегменты обнаружены.
9. Находим маркер «прочитан вслух».
10. Завершаем фазу сегментации.

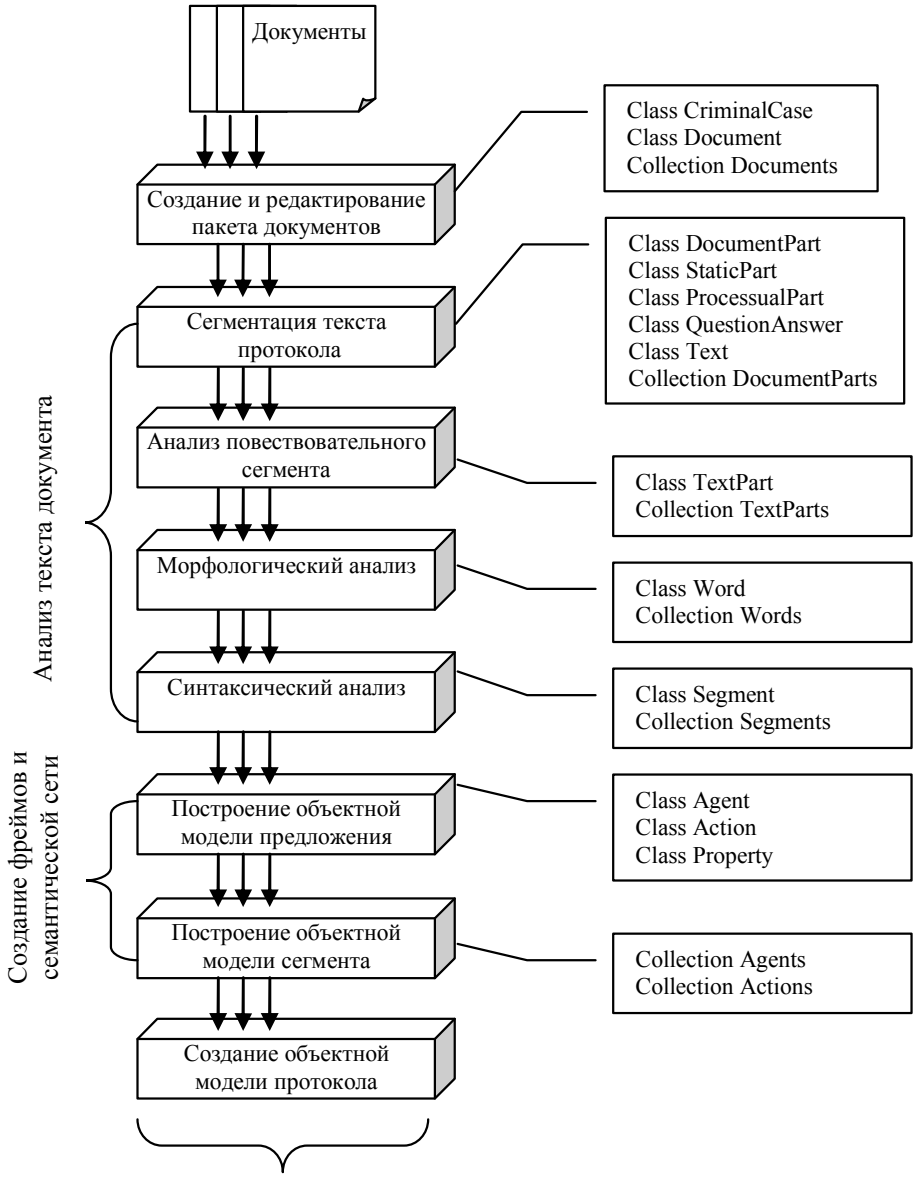

Рис. 3. Схема построения информационной модели протокола допроса

Таблица 2

Сущности, определяемые в процессе анализа протокола

Этапы	Используемые классы	Коллекции	Алгоритмы и методы
Создание и редактирование пакета документов	CriminalCase, Document	Documents	Работа с файлами и файловой системой
Сегментация текста протокола	DocumentPart, StaticPart, ProcessualPart, QuestionAnswer, Text	DocumentParts	Поиск на основе шаблонов и маркеров
Анализ повествовательного сегмента	TextPart	TextParts	Выделение предложений и определение его синтаксических характеристик
Морфологический анализ предложения	Word	Words	Определение морфологических характеристик слова и его нормальной формы
Синтаксический анализ предложения	Word	Words	Выделение словосочетаний на основе их морфологических характеристик
Построение объектной модели предложения	Agent, Action, Property, Fact, Event	Actions, Agents, AgentProperty, ActionProperty	Сворачивание словосочетаний за счет выделения ключевого слова и его характеристик. Построение формы предложения (P`)
Построение объектной модели повествовательного сегмента	Agent, Action, Property, Fact, Event	Actions, Agents, AgentProperty, ActionProperty	Построение семантической сети фреймов
Создание объектной модели протокола и обобщенной модели дела	Agent, Action, Property, Fact, Event	Actions, Agents, AgentProperty, ActionProperty	Построение композиции семантических сетей

Анализ повествовательного сегмента. На вход подается текст повествовательного сегмента. На выходе модуля получим размеченный текст с выделенными в нем предложениями. При этом в тексте выделяются вопросительные, восклицательные предложения и прямая речь.

Морфологический анализ предложения. На вход поступает выделенное на предыдущем этапе предложение. На выходе модуля получим размеченное предложение, в котором каждому слову приписаны его морфологические характеристики (часть речи, род, склонение, падеж, время и т. д.). Анализ каждого отдельного слова производится посредством:

– синтаксического анализа для выделения самого слова;

– нахождения по слову его морфологических характеристик с использованием базы тезауруса компании Dialing (http://www.aot.ru), построенного на основе словаря А.А. Зализняка, словарей имен собственных и географических названий в виде com-объектов;

– выделения даты и времени на базе морфологических признаков;

– выделения имен собственных (названия предприятий, географических объектов и т. д.);

– определения анафорических ссылок («он», «который» и др.) и обобщающих кореферентов («семьянин», «гражданин» и др.), а также сопоставления их соответствующим словам (лицам).

Синтаксический анализ предложения. По морфологическим данным, полученным на предыдущем этапе, в предложении выделяются связные словосочетания (сегменты).

Сегментация – это процесс группировки элементов текста, задающий однозначное отображение множества элементов на множество сегментов текста [4].

На основе свойства проективности русского языка [5] можно утверждать, что сегменты текста не пересекаются, но могут полностью содержаться в других сегментах. Любое сколь угодно сложное предложение можно разбить на сегменты. Если эти сегменты будут удовлетворять условиям синтаксической связности элементов и полноты, то анализ предложения сводится к анализу отдельных сегментов и установлению связей между ними.

Любое сложное предложение можно разбить на простые, которые содержат не более одного подлежащего или не более одного сказуемого [6]. Подлежащее и сказуемое образуют грамматическую основу предложения, а все остальные члены предложения выступают в роли подчинения грамматической основе.

Каждый сегмент имеет ключевое слово, которое несет в себе всю семантическую нагрузку сегмента, и второстепенные слова, которые дополнительно характеризуют ключевое слово. Тип ключевого слова определяется именным сегментом, именным сегментом с числительным, сегментом-датой, глагольным сегментом, сегментом с причастным оборотом, сегментом с деепричастным оборотом.

Итак, все рассматриваемые предложения обладают свойством проективности, т. е. сегменты, согласованные по морфологическим признакам, не пересекаются и, в крайнем случае, могут содержать в себе целиком другие сегменты:

$$P = < s_1, s_2, ..., s_n >,$$

где n – количество сегментов s_i в предложении P.

Так как один сегмент может содержать в себе набор сегментов, то все предложение можно представить в виде дерева, корнем которого выступает само предложение, а узлами – сегменты (рис. 4).

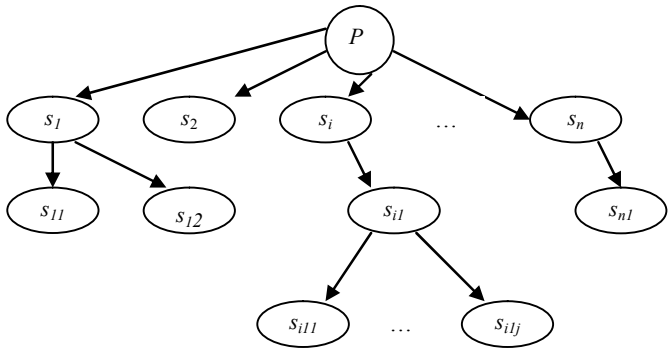

Рис. 4. Схема сегментов предложения

При разборе предложения выделяются главные члены (подлежащее и сказуемое) и группы определения, обстоятельства и дополнения.

Определение. *Простым предложением* назовем синтаксическую единицу, образованную одной синтаксической связью между подлежащим и сказуемым или одним главным членом [7].

Лемма 1. Для простого предложения сказуемое выделяется в отдельный сегмент и, как правило, состоит из одного слова.

Лемма 2. Если сегмент s_i содержит в себе вложенные сегменты, то можно построить сегмент s_i`, который будет листом дерева.

На основании леммы 2 дерево сегментов предложения P можно свернуть в дерево с глубиной 1 и таким образом получить новое предложение P`:

$$P \to P`,$$

где $P`=< s_1^{'}, s_2^{'}, ..., s_n^{'} >$ будет состоять из простых сегментов.

Поскольку каждый сегмент имеет ключевое слово, последнее преобразованное предложение можно свернуть еще раз, оставив в каждом из сегментов только ключевые слова в нормальной форме:

$$P``=< p,c,d\_p,d\_c,o,r >,$$

где p – подлежащее, c – сказуемое, d_p – дополнение к подлежащему, d_c – дополнение к сказуемому, o – обстоятельство, r – определение.

В процессе свертки и анализа сложного предложения будем выделять сущности, формируя тем самым базу объектов контекста.

Все слова естественного языка на основании морфологических характеристик классифицируются по частям речи, которые определяют не только синтаксические особенности использования слова в предложении, но и его семантическую роль. С целью построения более реалистичной модели контента проведем кластеризацию слов, выделив для каждой части речи набор кластеров, характеризующихся определенными свойствами: глаголы движения («пошел»), ощущения («почувствовал»), действия («ударил») и др.; существительные как географические объекты («город»), предметы («стол»), субъекты («девушка») и др. Таким образом в процессе анализа предложения согласно базе кластеров выделим сущности с заведомо определенными свойствами. В частности, если речь идет о некотором предмете, то в нем можно выделить цвет, вес, габариты и другие параметры. В рамках словосочетаний определим свойства выделенных сущностей. Например, в словосочетании «красная сумка» выделяем существительное «сумка». Оно относится к кластеру «предмет», а прилагательное «красная» определяет свойство этого предмета – «цвет».

Построение объектной модели предложения. В результате свертки получим набор ссылок на сущности с уже определенными свойствами. Моделью представления таких сущностей предлагается выбрать фреймы. Фрейм представляет собой не одну конкретную ситуацию (предмет или состояние), а наиболее характерные, основные моменты ряда близких ситуаций, принадлежащих одному классу. При этом составляющие фрейма могут быть определены или не определены (так называемые терминалы), что позволяет наиболее точно представлять модели реального мира. Кроме того, каждое предложение обычно описывает одно или несколько событий либо один или несколько фактов. Такие события и факты будем отражать в виде семантической сети фреймов. Смысловой портрет текста – это набор событий и фактов, упорядоченных во времени, а объектная модель предложения – это семантическая сеть связных фреймов.

Каждое событие характеризуется местом и временем, а также набором объектов и субъектов, принимающих участие в событии.

На базе выделенных событий, фактов, субъектов и объектов события строим объектную модель предложения.

Построение объектной модели повествовательного сегмента выполняется на основе согласования наборов объектных моделей предложений некоторого сегмента.

Создание объектной модели протокола и обобщенной модели дела осуществляется посредством согласования объектных моделей сегментов.

3. Диаграмма классов информационной модели

Информационная модель протокола допроса на каждом этапе описывается соответствующим набором классов, которые можно представить в виде диаграмм и модели их взаимодействия (рис. 5–8).

Опишем основные сущности диаграммы, представленной на рис. 5:

Collection – коллекции объектов некоторого типа, которые описываются единым интерфейсом *Collection.* Для коллекции допустимы операции добавления в коллекцию, удаления элемента из коллекции, подсчета количества элементов в коллекции до получения элемента коллекции по его номеру.

CriminalCase – класс, который описывает рассматриваемое дело, содержит в себе набор элементов (документы), а также специальную информацию о криминальном деле.

Documents – коллекция документов дела, состоящая из элементов типа *Document.*

Document – это отдельный документ дела, содержащий информацию о создателе, дате, названии и др.

Объекты диаграммы классов создаются на этапе формирования криминального дела. На вход этапа формирования криминального дела подаются документы, которые классифицируются и упорядочиваются.

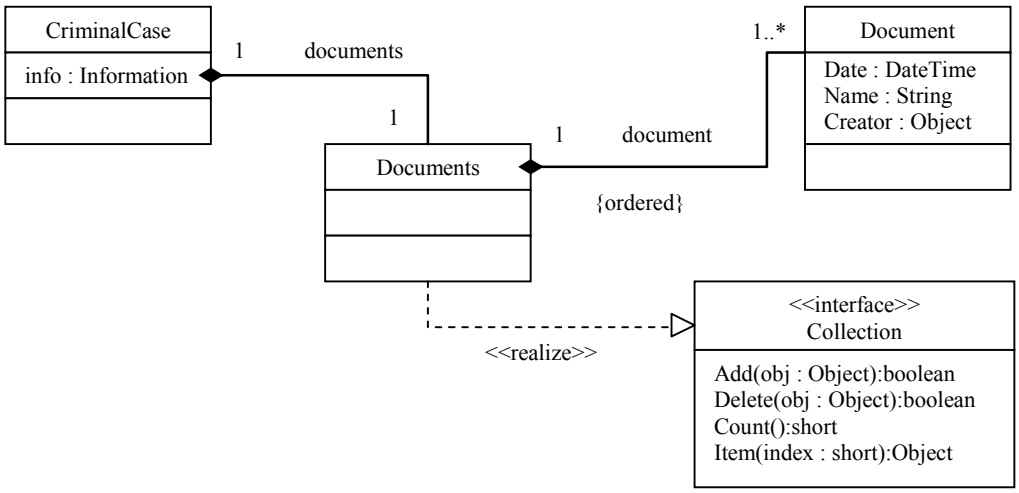

Рис. 5. Диаграмма классов формирования криминального дела

Каждый документ состоит из нескольких частей, объединенных в *DocumentParts* – коллекцию сегментов документа (рис. 6).

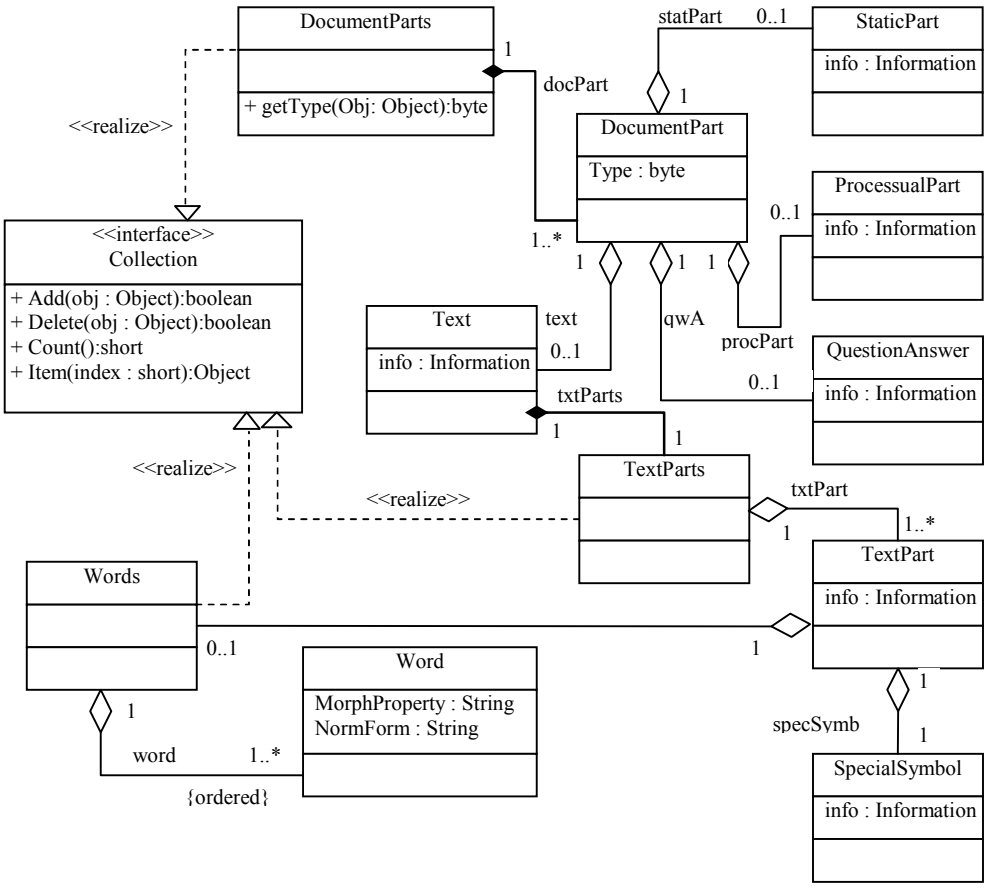

Рис. 6. Диаграмма классов анализа документов дела

DocumentParts – коллекция, состоящая из элементов типа *DocumentPart.* Элементы *DocumentPart* могут быть типа *QuestionAnswer* (вопросно-ответная часть), *Text* (повествовательная часть), *StaticPart* (постоянная часть) и *ProcessualPart* (процессуальная часть). Каждый тип части документа анализируется по собственному алгоритму.

Text – класс, который описывает повествовательную часть документа и содержит коллекцию нормализованных частей текста (простых предложений) *TextParts.*

TextParts – коллекция, состоящая из объектов типа *TextPart. TextPart* – часть текста, которая может быть отдельным словом или словосочетанием (коллекция *Words*) или специальным символом *SpecialSymbol.*

Words – коллекция слов и/или словосочетаний типа *Word*. Тип *Word* определяется набором морфологических характеристик, нормальной формой и т. д. Он представляет собой агента действия (*Agent*), само действие (*Action*) либо свойство действия или агента (*Property*) (рис. 7).

Agent характеризуется первоначальным, текущим и конечным состояниями, а также типом (человек, животное, здание-строение, географический объект и т. д.), образует коллекцию агентов (*Agents*).

Action характеризуется местом, временем и может быть либо событием (*Event*), либо фактом (*Fact*). Данные характеристики образуют коллекцию действий (*Actions*).

Property характеризуется типом и может быть свойством действия (*ActionProperty*) или свойством объекта (*AgentProperty*). Хранится в коллекции свойств агента или свойств действия.

Все агенты объединены в коллекцию *Agents*, все действия – в коллекцию событий *Events*, а все свойства – в коллекцию *Property*.

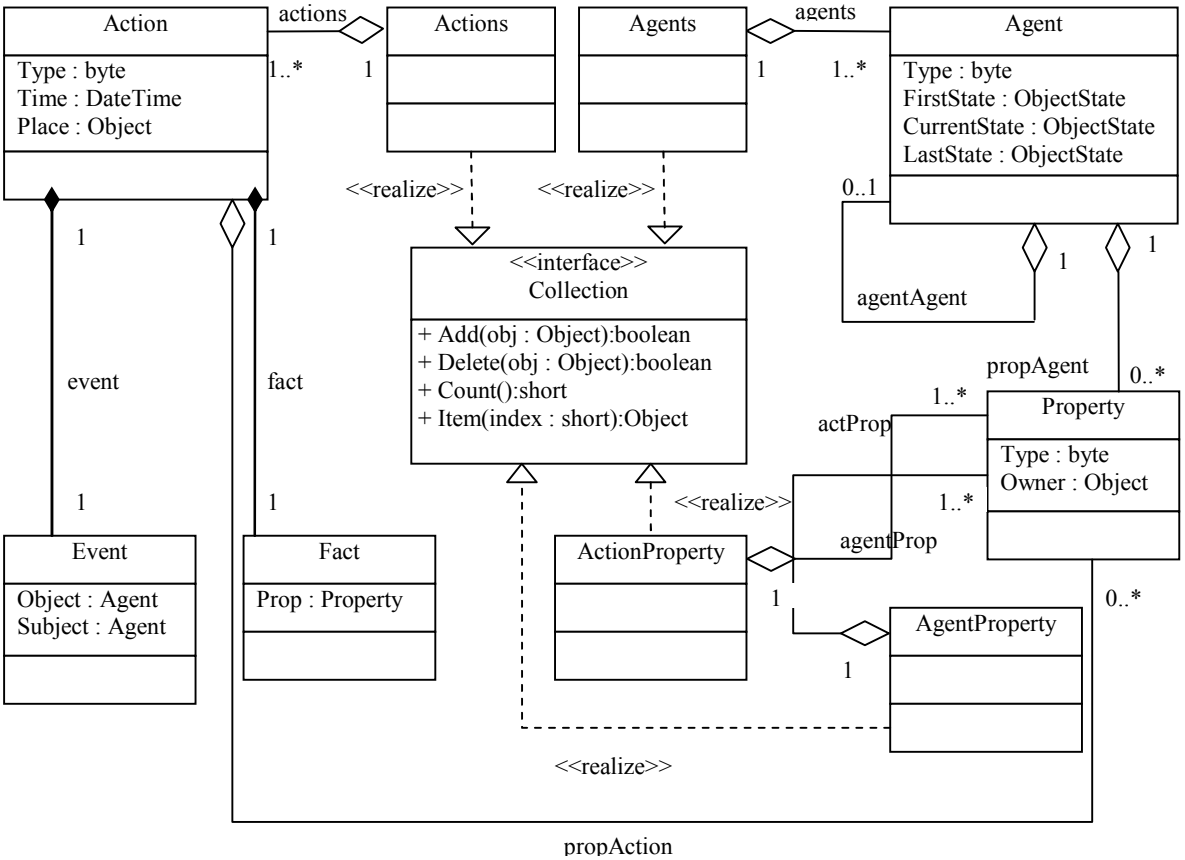

Рис. 7. Диаграмма классов формирования объектной модели

Для хранения состояния агента введен класс *ObjectState*, текущего состояния модели – *CurrentState* и протокола изменения состояния агентов – *ProtocolStateModify* (рис. 8).

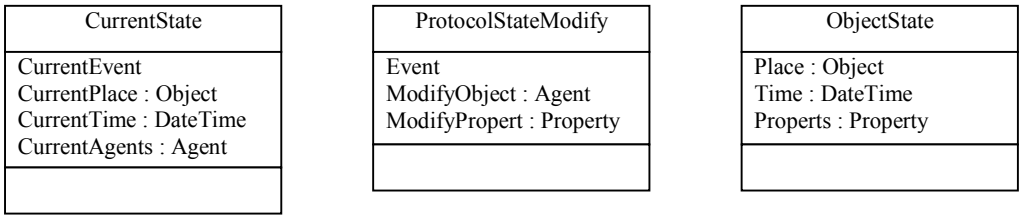

Рис. 8. Вспомогательные классы

Взаимодействие данных в описанных этапах показано на рис. 9.

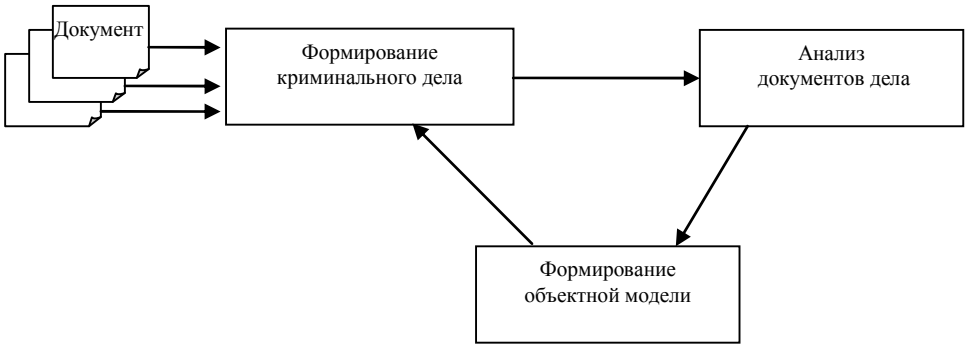

Рис. 9. Схема взаимодействия данных при построении информационной модели

На вход этапа формирования криминального дела подается набор текстовых документов, которые приобщаются к делу. Они классифицируются и упорядочиваются по времени поступления. Затем каждый из них проходит через этап анализа документов дела, в котором проводится синтаксический и морфологический анализы, выделяются сегменты и определяются данные для этапа формирования объектной (информационной) модели. На завершающем этапе формирования объектной модели определяются сущности: объекты, субъекты, факты, события и их свойства.

Заключение

Построенная объектная модель позволяет следователю принимать решения, руководствуясь проведенным анализом. Вместе с тем система, построенная на базе модели, может выступать и как своеобразный советчик для эксперта, позволяющий производить корректировку, настройку и модификацию проблемных участков.

Автоматизированная система на основании документов, составленных в ходе следственного процесса, позволяет упростить работу следователя по принятию решений и уменьшить затраты времени на рутинный процесс по формированию различного рода документации и ее анализу. Естественно, профессиональный следователь согласует полученные из системы данные с собственным опытом принятия решений, последними теоретическими исследованиями в этой области и новыми теоретико-техническими средствами, позволяющими оптимизировать процесс принятия решений.

Предложенная автоматизированная система построения информационной модели создана на платформе разработки приложений .Net Framework 3.5 средствами языка программирования C#.

Список литературы

1. Сборник образцов уголовно-процессуальных документов с комментариями. Возбуждение уголовного дела и предварительное расследование : учеб.-практ. пособие / авт.-сост. Г.Н. Васильев и др. ; под рук. и науч. ред. проф. М.А. Шостака. – Минск : Амалфея, 2006. – 704 с.

2. Малый энциклопедический словарь. – Т. 3 : репринтное воспроизведение издания Ф.А. Брокгауза, И.А. Ефрона. – М. : Терра – Книжный клуб, 1997. – 560 с.

3. Зорин, Г.А. Многовариантные программы допросов: технологии построения и применения / Г.А. Зорин. – М. : Изд-во деловой и учебной литературы, 2005. – 336 с.

4. Невзорова, О.А. Алгоритмы сегментации предложений на простые составляющие / О.А. Невзорова, М.П. Сергеев // Тр. Междунар. семинара «Диалог'2000» по компьютерной лингвистике и ее приложениям [Электронный ресурс]. – Режим доступа : http://www.dialog-21.ru/materials/archive.asp?id=6528&y=2000&vol=6078. – Дата доступа : 15.04.2009.

5. Гладкий, А.В. Синтаксические структуры естественного языка / А.В. Гладкий – М. : Изд-во ЛКИ, 2007. – 142 с.

6. Ельчанинова, Н.Б. Проблемы извлечения знаний из текстов нормативно-правовых актов и их структурирование / Н.Б. Ельчанинова // Теория права и правовая информатика. – Вып. 2. – Ростов-на-Дону : Изд-во РЮИ, 2002. – С. 27–45.

7. Розенталь, Д.Э. Справочник по русскому языку. Словарь лингвистических терминов / Д.Э. Розенталь, М.А. Теленкова. – М. : Оникс, 2008. – 624 с.

8. Троелсен, Э. Язык программирования C# 2005 и платформа .NET 2.0 / Э. Троелсен. – Киев : Вильямс, 2007. – 1168 с.

Поступила 31.07.09

[1]*Белорусский государственный университет,*
Минск, пр. Независимости, 4
e-mail: bouza@bsu.by

[2]*Гродненский государственный*
университет им. Я. Купалы,
Гродно, ул. Ожешко, 22
e-mail: nvdeeva@gmail.com

M.K. Bouza, N.V. Deeva

INFORMATION MODEL OF THE OFFENCE INVESTIGATION PROCESS

The paper addresses the problem of building an information model for processing criminal investigations, in the context of management of interrogation transcript reports. The transcript text is analyzed, and the main sections are highlighted along with the corresponding algorithms. An algorithm for building a generalized object model for a criminal case is proposed, and the data classes along with the methods of their presentation are considered.