Интеллектуальный анализ данных и распознавание образов

От графического образа к универсальному представлению Формы документа

В. В. Арлазаров, В. Е. Кривцов, Д. В. Полевой, Д. Г. Слугин, И. М. Янишевский

Аннотация. Рассматривается проблема моделирования формы структурированного документа. Предложена универсальная модель формы представления документа, позволяющая связывать и описывать основные процессы ввода/вывода бумажных и электронных документов. Дано описание каждого элемента конструкции форм. Особое внимание уделено отражению преимущества рассматриваемой концепции перед известными подходами.

Ключевые слова: *концепция Формы, универсальное представление документа, модель содержания, модель визуализации, модель взаимодействия, секционная модель.*

Введение

В современном мире ежедневно вводятся с бумаги, заполняются на экранах компьютеров, обрабатываются различным образом и выводятся на бумагу миллиарды различных форм документов: почтовых карточек, платежных поручений, таможенных или налоговых деклараций, банковских чеков, бюллетеней для голосования, разного рода бумажных и электронных анкет, заказов на товары или услуги в электронных магазинах, разных отчетов, деклараций и множество других видов. Сотни тысяч операторов выполняют рутинную последовательность действий при вводе форм: бросают взгляд на очередную страницу, находят текст заполнения и набирают его на клавиатуре. Активно используются тысячи различных систем электронного документооборота, базирующихся на понятии «форма»; эти системы применяются практически во всех сферах деятельности. Альтернативой ручному вводу являются технологии автоматизированного ввода форм, что позволяет существенно сократить время ввода данных с форм и уменьшить расходы на персонал.

В данный момент основной тенденцией стало объединение систем бумажного и электронного документооборота в единые комплексные системы, в которых одновременно присутствуют как бумажные, так и электронных документы. В них интегрируются системы ввода/вывода бумажных документов и системы ввода/вывода электронных документов. Примером таких систем могут быть системы учета и обработки налоговых деклараций, которые можно подавать как в бумажном виде, так и в виде электронного pdf файла или заполнить на сайте в сети Интернет. Форма анкеты декларации, размещенная в глобальной сети, разосланная по электронной почте и напечатанная на бумаге, по сути, это одинаковые формы, содержащие одни и те же вопросы, их отличие заключено в способе представления и частично в способе взаимодействия с пользователем. Без единой модели формы необходимо будет создать несколько описаний электронных докумен-

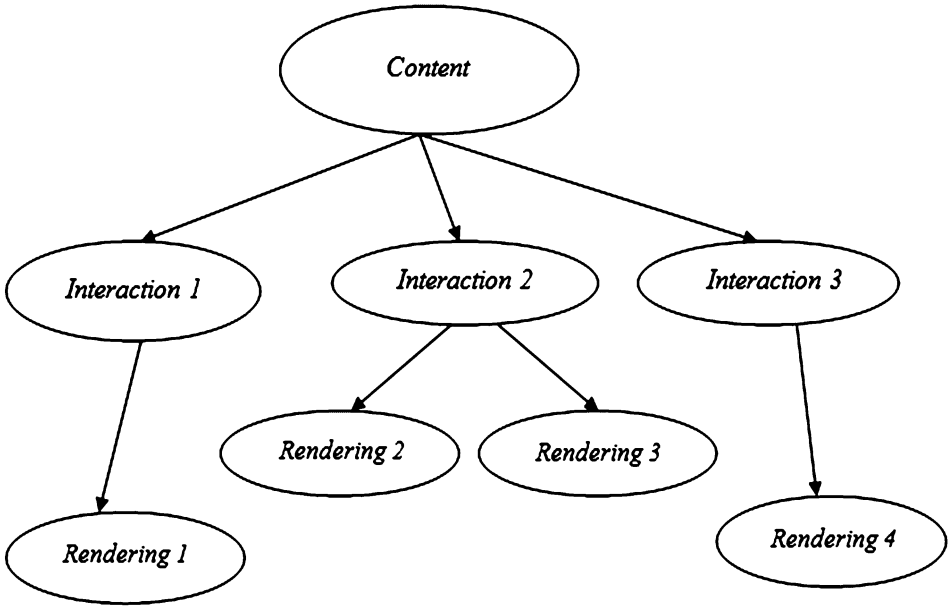


Рис. 1. Пример схемы Формы

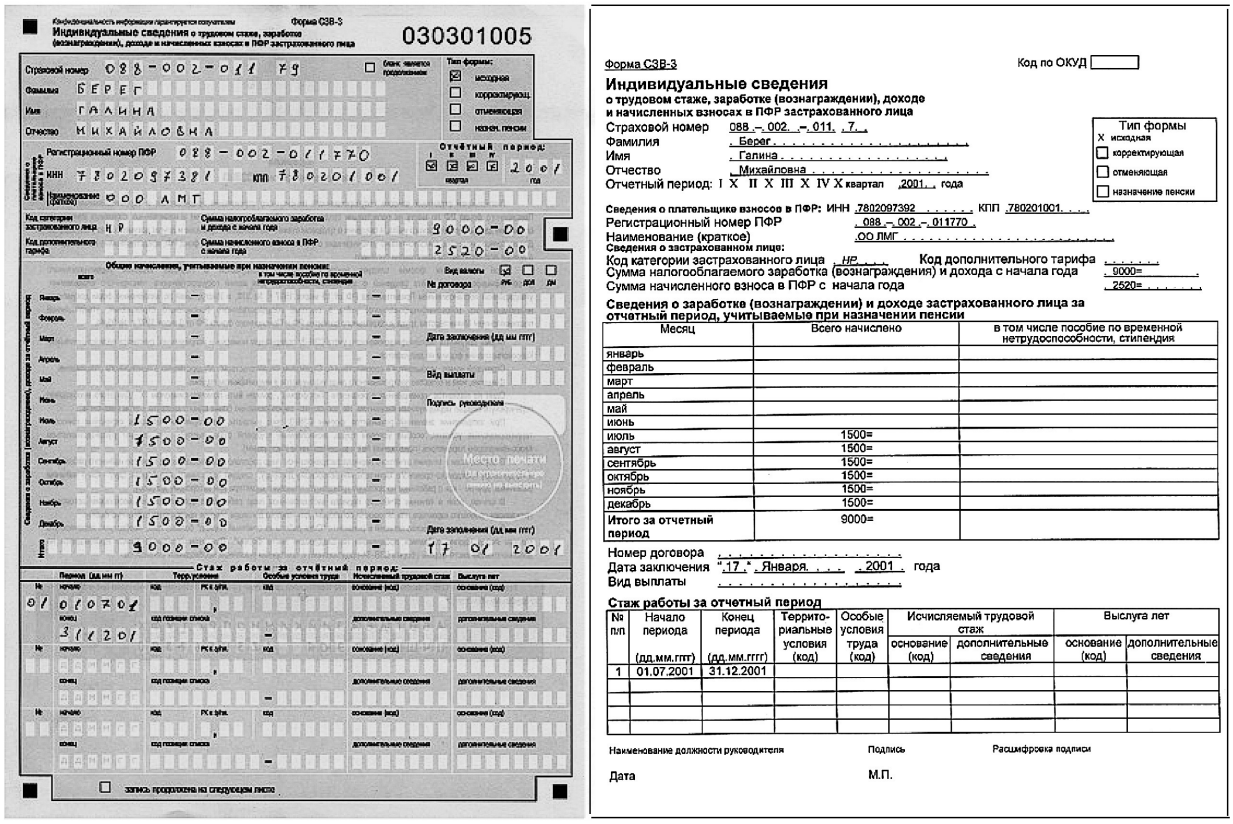
тов — для глобальной сети, электронных файлов и бумажных документов, при этом большая часть спецификаций (описание данных, правила проверки и заполнения и прочее) будет дублироваться. При этом необходимо будет воспользоваться несколькими различными системами описания формы и языками программирования для создания этих форм. После чего нужно реализовать обработку различных заполнений этих форм, используя разные средства разработки. Построение единого подхода к форме в разных ее проявлениях и создание модели формы является актуальной задачей при построении комплексных систем документооборота.

Предметом данной статьи является анализ и выявление общности, обеспечивающей конструктивную основу для решения задач ввода/вывода структурированных документов для систем документооборота стандартных форм. Целью статьи является построение концептуальной модели формы, которая бы позволяла органично связать и описать основные процессы ввода/вывода структурированных документов.

Предложенный подход состоит, прежде всего, в разработке универсальной модели формы структурированного документа, используемой для различных задач и абстрагированной от конкретных методов обработки, в отличие от существующих подходов, ориентированных на представление либо экранных, либо бумажных форм. Впервые модель объединяет процессы ввода/вывода вне зависимости от того, экранная или бумажная форма используется в них. Независимость модели от особенностей конкретных методов обработки обеспечивает ее открытость для разработки и подключения новых методов, расширение классов обрабатываемых документов в рамках предложенной концепции.

1. Определение Формы

Слово Форма определяется по словарю Webster's New Collegiate Dictionary как распечатанный на машинке или типографский документ с пустыми местами для заполнения. В других работах Форма определяется как логический образ документа. В данном контексте Форма определяется как информационный объект, представляющий собой логический образ (описание) класса информационных объектов, который позволяет порождать информационные объекты этого класса различными способами, гарантируя структурную, внутреннюю и внешнюю логическую непротиворечивость, и обеспечивает взаимодействие информационных объектов и объектов реального мира (например, человека), преобразование объектов данного класса в объекты другого класса, в том числе и с участием человека, и отображение информационных объектов в объекты реального мира и обратно. Чтобы не было путаницы, будем называть информационный объект и писать с большой буквы Форма, а заполнение называть экземпляром формы и писать просто форма. Таким образом, можно сказать, что Форма *FORM f* — составной информационный объект, однозначно определяющий класс экземпляров *IMPLEMENTATION Im* формы.



Форма СЗВ-3

Код по ОКУД

Индивидуальные сведения
о трудовом стаже, заработке (вознаграждении), доходе и начисленных взносах в ПФР застрахованного лица

Страховой номер 088-002-011 79
Фамилия Берег
Имя Галина
Отчество Михайловна
Отчетный период: I X II X III X IV X квартал 2001 года

Тип формы: X исходная; корректирующая; отменяющая; назначение пенсии

Сведения о плательщике взносов в ПФР: ИНН 7802097392 КПП 780201001
Регистрационный номер ПФР 088-002-011770
Наименование (краткое) ООО ЛМГ

Сведения о застрахованном лице:
Код категории застрахованного лица НР
Код дополнительного тарифа
Сумма налогооблагаемого заработка (вознаграждения) и дохода с начала года 9000=
Сумма начисленного взноса в ПФР с начала года 2520=

Сведения о заработке (вознаграждении) и доходе застрахованного лица за отчетный период, учитываемые при назначении пенсии

Месяц	Всего начислено	в том числе пособие по временной нетрудоспособности, стипендия
январь		
февраль		
март		
апрель		
май		
июнь		
июль	1500=	
август	1500=	
сентябрь	1500=	
октябрь	1500=	
ноябрь	1500=	
декабрь	1500=	
Итого за отчетный период	9000=	

Номер договора
Дата заключения "17" Января 2001 года
Вид выплаты

Стаж работы за отчетный период

№ п/п	Начало периода (дд.мм.гггг)	Конец периода (дд.мм.гггг)	Территориальные условия (код)	Особые условия труда (код)	Исчисляемый трудовой стаж: основание (код)	Исчисляемый трудовой стаж: дополнительные сведения	Выслуга лет: основание (код)	Выслуга лет: дополнительные сведения
1	01.07.2001	31.12.2001						

Наименование должности руководителя | Подпись | Расшифровка подписи

Дата М.П.

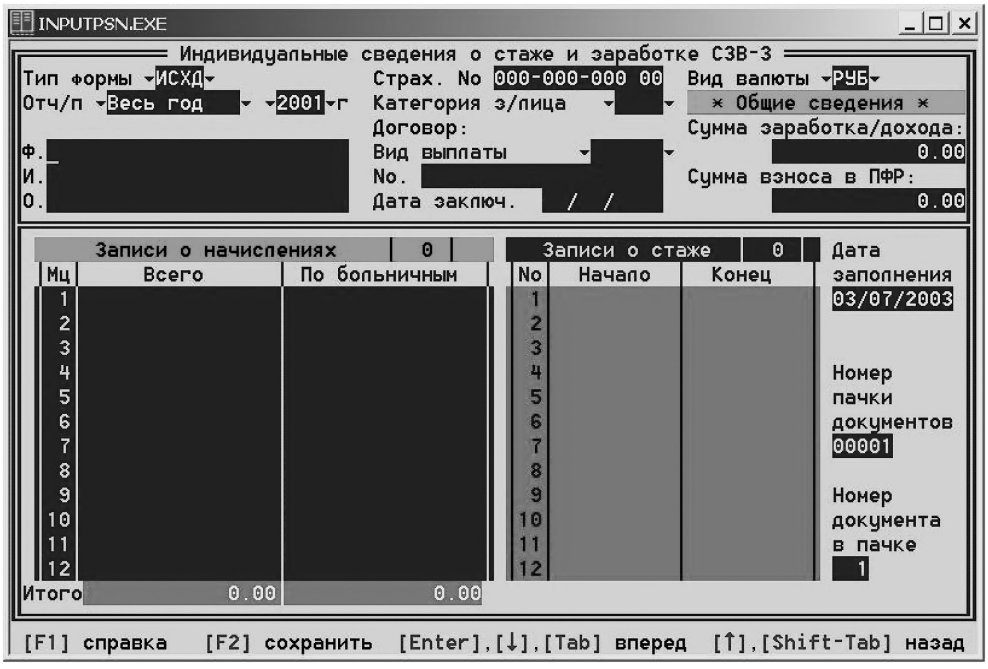


Рис. 2. Пример одной и той же формы СЗВ-3 для ручного, печатного и экранного заполнения

Одно из основных положений данной работы состоит в следующем: Форма разделена на три модели:

- *CONTENT c* — модель содержания (схема базы данных);
- *RENDERING r* — модель визуализации (способ ввода/вывода);
- *INTERACTION i* — модель взаимодействия с пользователем (описание человеко-машинной процедуры преобразования данных реального мира в структуры описания базы данных).

Такое разделение объясняется следующими причинами. Разделение содержания и образа информационного объекта типично и вызвано тем, что один и тот же объект в различных процессах может иметь различный образ, но содержательно оставаться неизменным. Также одна и та же форма в типографском исполнении, на принтере и на мониторе почти всегда выглядит по-разному.

Основные принципы синтеза Формы формулируются следующим образом. В любой Форме обязательно присутствует модель содержания и хотя бы одна модель взаимодействия с пользователем, визуальная модель может, как присутствовать, так и отсутствовать, кроме того, в одной Форме может присутствовать несколько моделей взаимодействия с пользователем и несколько визуальных моделей. Модель содержания по определению единственна.

2. Концепция формы

Данная концепция формы разрабатывалась для применения в следующих процессах:

- ввод информации с дисплея в формализованном структурированном виде,
- редактирование информации в формализованном структурированном виде на дисплее,
- ввод структурированных документов с бумаги,
- печать структурированных документов на различных устройствах,
- создание электронных аналогов Формы в популярных форматах.

Разрабатываемая здесь модель формы, как всякая более четко структурированная по сравнению с существующими моделями обладает рядом преимуществ. Во-первых, она задает методику, по которой быстрее строятся конкретные системы. Во-вторых, она дает основу для выделения подсистем, которые могут быть использованы в различных процессах. Так, имея четыре стандартные основные подсистемы: «Загрузка/Выгрузка данных», «Построения экранной формы и взаимодействия с пользователем», «Распознавания формы», «Идентификация формы», можно построить процессы ввода, модификации, печати и распознавания форм, практически исключив написание операторов неспецифических для конкретных задач.

В-третьих, данная модель позволяет стандартизировать целый ряд интерфейсов, благодаря чему многие существующие программные средства могут подключаться, что называется в «стык».

Исходя из вышесказанного, можно построить логические модели процессов, в которых будет использована модель.

Ввод и модификация данных по Форме. Основной задачей процесса является предоставление пользователю быстрого и удобного для него способа создания и модификации информационного объекта. Процесс состоит из выгрузки данных/загрузки данных из объекта и взаимодействия с пользователем.

Печать Форм. Основной задачей процесса является осуществление проекции информационного объекта на бумагу, т. е. преобразование объекта информационного мира в документ реального мира.

Распознавание Форм. Распознавание Форм является одним из самых сложных процессов обработки. Основной задачей процесса является преобразование документа реального мира в информационный объект.

3. Модель содержания

Назначение: задание модели данных Формы, определение ограничений и правил заполнения, настройка и задание источника данных, спецификация способа отправки и получателя данных. Таким образом, основные задачи, решаемые моделью содержания, — структурирование, проверка данных и сохранение/получение данных. С точки зрения концепции модель содержания Формы есть дерево именованных объектов данных.

4. Модель визуализации

Модель визуализации описывает образ (электронный и бумажный) информационного объекта. Назначение — обеспечение процессов визуализации Формы и процессов идентификации и распознавания Формы. Теоретические основы построения модели отображаемого на бумагу электронного документа и классификация основных приемов визуализации сложноструктурированных данных представлены в работах Н. Е. Емельянова и его учеников. В рамках данной теории под документом понимается структурированный текст как совокупность

взаимосвязанных семантических блоков (некоторых фрагментов документа, выделенных по смысловому содержанию). Для решения этих задач может использоваться либо один из стандартных языков описания визуализации/распознавания XSL-FO и другие, либо свой язык визуализации/распознавания. Модель только задает общие синтаксические и грамматические принципы создания/использования таких языков. Модели визуализации и распознавания и языки, их описывающие, разрабатываются уже достаточно длительное время (печати — более 30–40 лет, вывода на экран дисплея более 25–30 лет, электронных форм — более 15 лет и распознавания более 25 лет) и представлены в большом количестве работ, список которых приведен в конце статьи.

5. Модель взаимодействия

Модель взаимодействия занимает промежуточное положение между моделью содержания и моделями визуализации и описывает схемы взаимодействия «пользователя» и данных Формы. В слоях этой модели содержится специфика процессов, в которых принимает участие Форма, и необходимые дополнительные описания и данные для обеспечения процесса обработки и целостности данных. Необходимость этой модели вызвана серьезными различиями между процессами, в которых участвует Форма. Например, в системе генерации отчетов содержимое информационного объекта преобразуется сначала в строковый вид (принтер понимает только строки), а только потом уже выводиться, при этом часто на основе анализа данных вычисляются различные новые поля Формы, которые отсутствуют как в информационном объекте, так и при заполнении этой же Формы на экране. Процесс распознавания требует значительного числа специфических настроек, актуальных только для него (например, алгоритм распознавания), и не всегда может гарантировать тип получаемых данных (на выходе у систем распознавания поля — строка символов). При этом модель содержания у всех этих процессов одна, да и модель визуализации у процессов генерации отчетов и у системы распознавания тоже может быть одна. Модель специфицирует события самой Формы (например, окончание ввода в поле) и процессов обработки (например, начало и окончание обработки формы), а также способ описания реакции на их возникновения. Задаются базовый набор событий и правила создания новых типов событий и их свойств (содержится в описании слоя «событийная модель»).

Фактически в слоях этой модели описывается способ представления данного в текстовом виде, способ взаимодействия с пользователем и содержательная разница между процессами, в которых может использоваться Форма. Рассмотрим модель взаимодействия на примере секционной модели Формы.

5.1. Секционная модель Формы

Назначение: описание способа представления документа в виде дерева текстовых (значения полей) и бинарных данных (например, картинок). Левый обход — порядок чтения. Уровни декомпозиции Формы:

- *Секция SECTION s* — элемент или полное поддерево Формы. Содержит текст, непосредственно относящийся к секции, и позволяет перейти к «дочерним» и «братским» секциям и содержит ссылку на схему секции.
- *Абзац PARAGRAPH p* — секция, не имеющая дочерних секций, т. е. следующий уровень декомпозиции — слова и буквы.

Модель представляет из себя ориентированный граф, задающий описание сегментации в дерево секций класса Форм. Вершинами этого графа являются *схемы секций*, а ребрами — *схемы переходников к следующей секции*. *Схема секции SCHEMA SC*(*s*, *p*) содержит описание секции Формы и задает тип секции. Основные атрибуты: UUID схемы секции, тип секции, заголовок и само описание: тип, формат и т. д. Схема секции может быть простой — описывает терминальную секцию (абзац) — или составной — описывает комбинацию секций и абзацев. Модель составной секции, как и модель Формы, представляет собой ориентированный граф подсекций. Разделение на секции производится исходя из логической целостности ее данных и необходимости процесса, обрабатывающего данную Форму. Схема *переходника к следующей секции* задает, секции какого типа могут следовать далее, начиная с данного места. Для этого используется атрибут, допустимое число переходов, значение которого может задаваться как фиксированное число, так и как некоторая функция от данных секции(й) и текущего состояния формы. Пример полученного графа и соответствующее ему разбиение на секции Формы СЗВ-3, предназначенной для распознавания, дается на рис. 3 и 4.

Секции бывают простыми (абзацы), и составными и описываются графами аналогично модели Формы, при этом их структура может быть как простой (секция номер 1, рис. 4), так и достаточно сложной (секция номер 6, рис. 4). Седьмая и восьмая секции этого документа являются опциональными, и их наличие зависит, в случае седьмой секции, от значений предыдущих секций, а в случае с восьмой

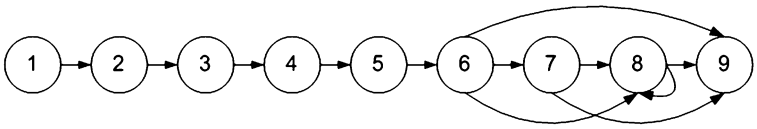


Рис. 3. Граф секций Формы СЗВ–3

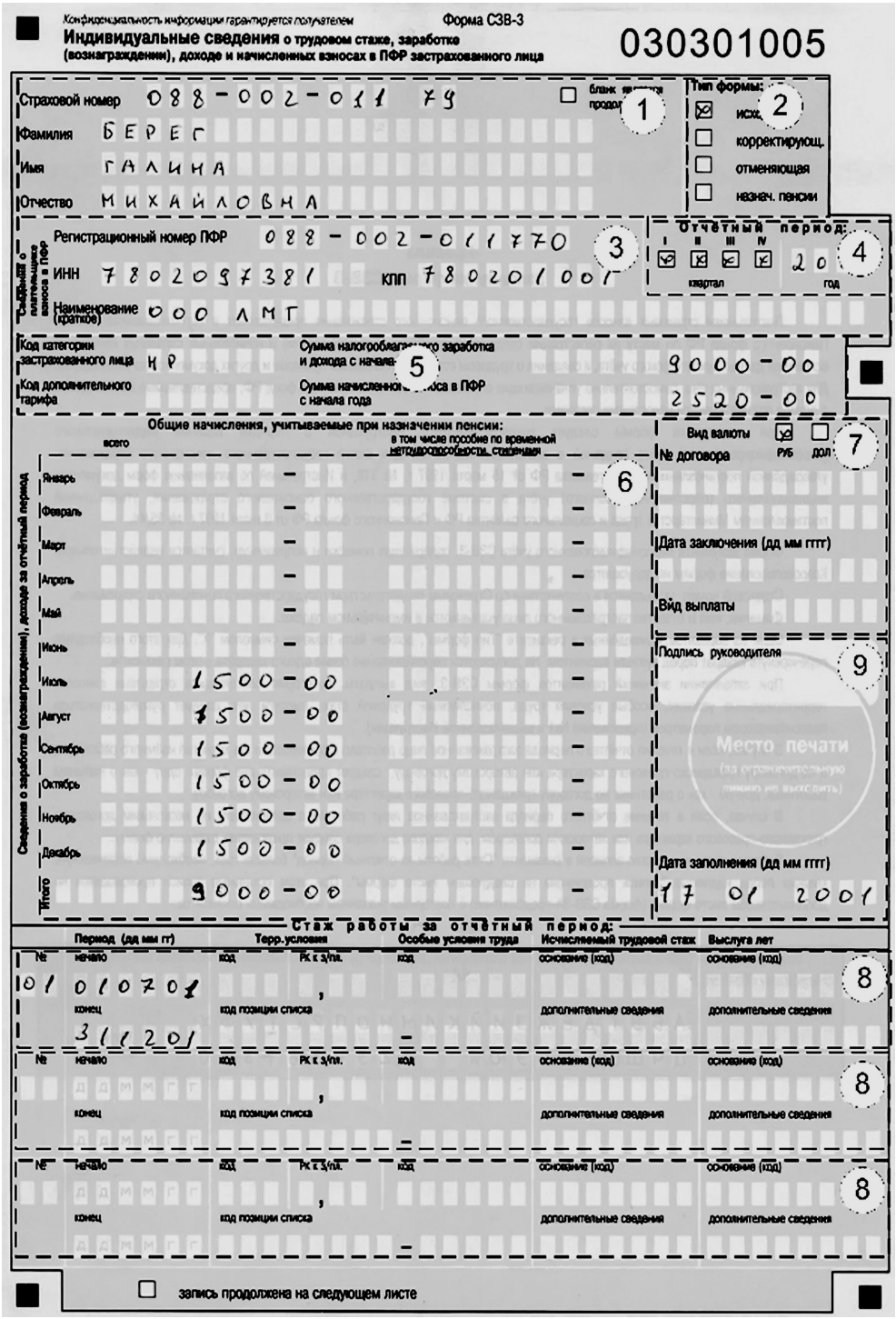


Рис. 4. Секции формы СЗВ-3

(наличие льгот по пенсии) — от внешнего мира. Кроме этого, в этой модели присутствуют переходники, задаваемые как числом (например, переход от первой секции ко второй возможен только один раз), так и функцией — число повторений восьмой секции не может быть больше, чем число заполненных периодов в шестой, и не может быть больше 12.

Необходимо отметить, что граф на рис. 3 описывает не только секционную модель СЗВ-3 в процессе распознавания, но и секционную модель печатной формы СЗВ-3 на рис. 2-2а, но отличен от секционной модели экранной формы СЗВ-3.

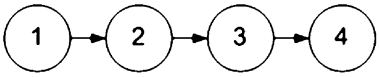


Рис. 5. Схема первой секции Формы СЗВ-3

Отметим также, что часть значений секций в секционной модели может вычисляться на основании предыдущих как некоторая функция от уже реализованного дерева секций. Правила, по которым это делается, задаются в описании схемы секции.

6. Базисные секции

Существуют типы секций, которые являются базисными для понятия «документ». Список и модели этих схем являются неотъемлемой частью концепции Формы, эти секции будем называть базисными. Например, тип секции «Сумма» рис. 7, состоящий из перечисления элементов — и элемента «Сумма» — , кроме того, сами элементы, все или их часть могут являться в свою очередь суммами. Примером такого типа секции служит шестая секция СЗВ-3 рис. 4, в этой секции есть две суммы — «Итого — общих начислений» и — «Итого — повременной нетрудоспособности...».

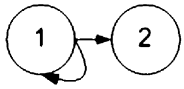


Рис. 7. Схема секции «Сумма»

Надо заранее отметить, что выбор базисных секций основывается не только на частоте их встречаемости, но и на основе семантического анализа документов реального мира. При анализе каждого

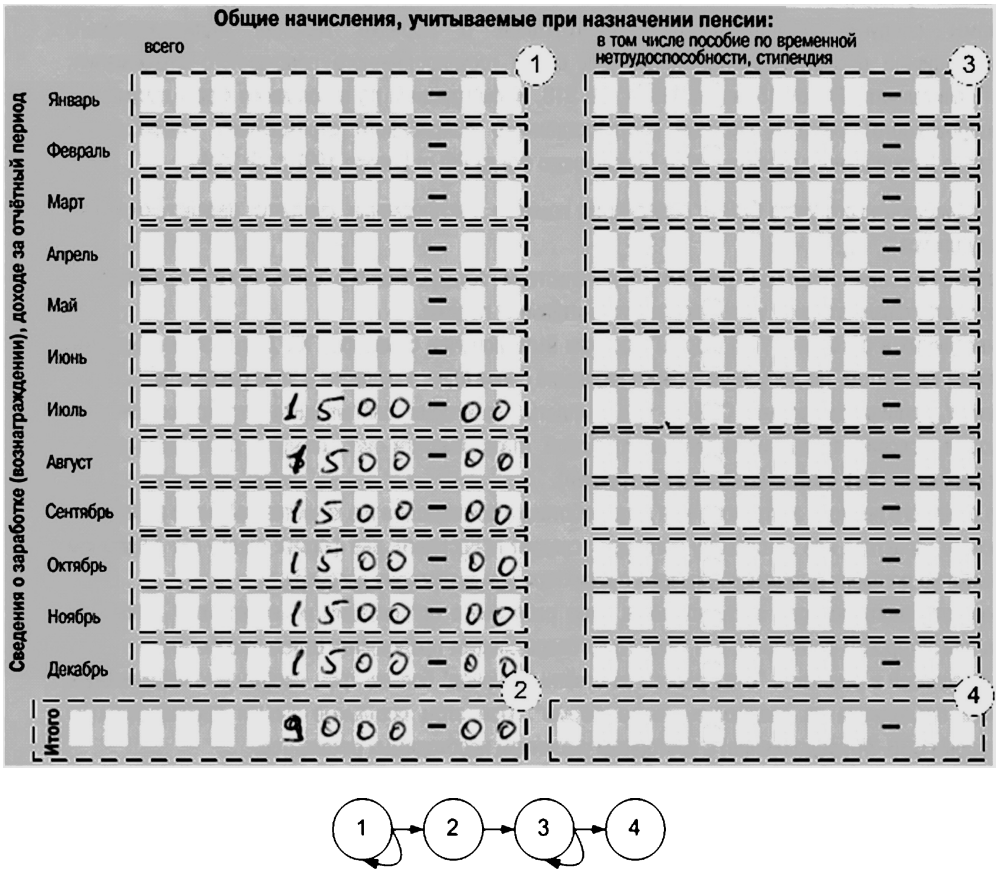


Рис. 6. Разбиение и схема шестой секции

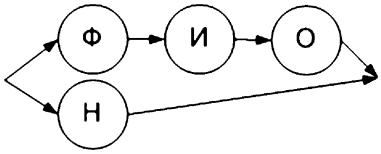


Рис. 8. Схема секции «ФИО с названием организации»

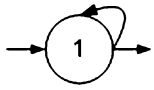


Рис. 9. Схема секции «Перечисление»

класса документов задается вопрос: «Какие семантические блоки являются неотъемлемыми для понимания содержания документа?». Например, для текстов исторических документов характерны блоки «Место», «Время», «Имя». Для деловых документов очень важны блоки «Адрес», «Имя», «Дата», «Сумма». А для научных документов понятие «Автор» — частный случай «Имени» — является практически обязательным и очень важным: «Дата» также является важной составляющей (нередко критической). Сутью почтовых документов — Почтовых карточек, Конвертов — являются понятия «Кому», «От кого» и «Адрес», при этом понятия «Кому» и «Кого» являются частными случаями «Имени», но при этом понятие «Время» в момент заполнения отсутствует. Большинство бухгалтерских документов, как впрочем и почти весь бухгалтерский учет, построены на основе понятий «Перечисление», «Сумма», «Баланс» и «Дата». При этом необходимо заметить, что понятие «Дата» есть сужение понятия «Время», а «Адрес» — более четко специфицированное понятие «Место». Исходя из синтеза и анализа таких блоков и был выстроен базисный набор секций. Более подробно с семантическим анализом документов можно ознакомиться в работах из списка литературы.

Разумеется, набор типовых секций не является «абсолютным знанием», а зависит от типа используемых документов. Если мы будем иметь дела с описаниями произведений живописи, то появится секция «размер» и т. п. Суть здесь не в полноте, а в наличии содержательных операций выделения, обработки контроля и оформления. Далее приведен некоторый список базисных секций, их схемы и комментарии.

1. «Имя» — тип секции, описывающий либо личные данные («фамилия», «имя», «отчество»), либо название организации, но не название документа — документ может и не иметь названия. Практически любой документ имеет упоминание либо личности, либо названия организации — документ содержит либо автора, либо он адресован кому-то, либо, если это «анонимка», то описывает кого-то в тексте. Очевидно, что документ по своей природе персонифицирован, и эта секция присутствует и во многом определяет его структуру, порядок, правила и особенности его обработки. Схема типа секции на рис. 8 состоит из обязательных терминальных секций «Фамилия» и «Имя», и опциональной «Отчество» (в случае описания) человека или из названия организации.

2. «Дата» — тип секции, описывающий понятие даты как секцию терминального типа. «Дата» — одно из наиболее фундаментальных и очень важных понятий мира документов, хотя есть примеры отсутствия даты на Форме — форма «Карточка почтовая» изначально не несет на себе ни даты, ни времени (дата и время появляются в процессе обработки (доставки) этой формы). Тип описывает дату в формате, принятом для деловых документов — день, месяц, год (4 символа) в числовой записи.

3. «Перечисление» — тип секции, описывающий секцию, состоящую из последовательности однородных подсекций. Этот тип секций весьма распространен во всех классах документов, а для бухгалтерских и отчетных документов является одним из основных и формирует понятие этого класса документов, также «перечисление» активно используется и в других классах документов. Примером перечисления может служить данный список базисных секций или чек кассового аппарата. Понятие «перечисления» используется и в качестве части другой базисной секции «Сумма». Схема типа представлена на рис. 9.

4. «Сумма» — тип секции, являющийся суммой перечисления, в том числе и перечисления других сумм, схема рис. 8. Типичные примеры класса секции «Сумма» разнообразные «Итого:» на различных документах типа «Счет...» (квитанция оплаты электричества, чек из магазина, счет в ресторане и т. д.). Причиной выделения этой секции в качестве базисной является то, что данная концепция разрабатывалась, в основном, для описания деловых документов с учетом пригодности и простоты расширения ее и для описания других классов документов, а большую долю деловых документов составляют бухгалтерские и отчетные формы, для которых, как уже говорилось выше, это понятие основополагающее.

5. «Адрес» — тип секции, описывающий адресную информацию, получен из понятия «Место» путем более четкой его спецификации. Разницу хорошо иллюстрирует следующий пример: на титульном листе этой работы последняя строка выглядит так: «г. Москва, 2004 год», фраза «г. Москва» задает место, где написана эта работа, но она не описывает адрес того, где она была написана, и по не-

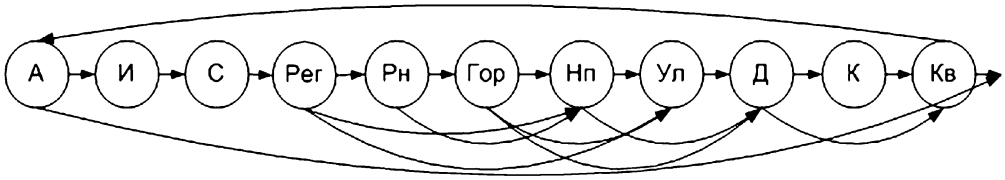


Рис. 10. Схема секции «Адрес»

му невозможно послать сообщения с критикой или поддержкой данной работы. Секция описывает адресную информацию в принятом в Российской Федерации виде. Схема представлена на рис. 10, где А – адрес, записанный по канонам русского языка, И – индекс, С – страна, Рег — регион РФ, Рн — район региона РФ, Гор — город, Нп — населенный пункт (село, деревня, ПГТ и т. д.), Ул. — улица (проспект, площадь и т. д.), Д — дом, К – корпус, Кв. — квартира. Адрес разделяется на два блока с одной общей вершиной — индекс. Первый блок — адрес, записанный по правилам русского языка, второй блок — представление адреса в стандартизированной форме из 10 полей, которая используется для его представления в информационных системах. При выводе информации происходит преобразование, при вводе информации — обратное, надо отметить, что задача обратного преобразования значительно сложнее прямого. Особенности задачи структурирования и способы ее решения описаны в работе [7]. Приведем пример того, как выглядят неструктурированный и структурированный адреса:

a) неструктурированный адрес: 117312 г. Москва проспект 60-Летия Октября дом 9, 604;
b) адрес, разложенный в структуру: 117312, РОССИЯ, МОСКВА Г,,,, 60-ЛЕТИЯ ОКТЯБРЯ ПР-К, 9,, 604.

Интересной особенностью данного примера является то, что «МОСКВА Г» играет роль региона, а не города, хотя и является городом.

6. «Реквизиты физического лица» — тип секции, описывающий идентификационные данные человека: паспортные или иные удостоверяющие данные, ИНН, номер социального и пенсионного страхования. Данная секция внесена в базис по двум причинам:

a) существует большой класс деловых документов — договоры, различные анкеты и другие официальные документы которые, во-первых, не являются содержательными документами без наличия этой секции, а во-вторых, в них необходима точная идентификация секции типа «Имени». Примером такого документа может служить документ паспорт — сам смысл, которого — предоставить часть данных этой секции;
b) данные этой секции в современном деловом мире являются практически неотъемлемой частью «Имени».

Выделение ее в отдельную секцию вызвано постоянно меняющимся форматом этой секции, тогда как тип секции «Имя» достаточно устойчив. Схема типа секции представлена на рис. 11, где С – серия, Н — номер, КВ — кем выдан, ДВ — дата выдачи, Нд — название документа, ИНН — идентификационный номер налогоплательщика, Соц — номер свидетельства социального страхования, ПФ — номер свидетельства пенсионного страхования. Подсекции Нд, ИНН, Соц, ПФ — являются опциональными.

7. «Реквизиты организации» — тип секции, описывающий идентификационные данные организации в соответствии с принятыми в Российской Федерации нормами. Схема секции на рис. 12, где: ИНН — идентификационный номер налогоплательщика организации, КПП — код причины постановки на учет в налоговом органе организации, Счет — счет организации в банке, БИК — банковский идентификационный код, СчБ — счет банка. Назначение и причины выделения в отдельный от «Имени» тип

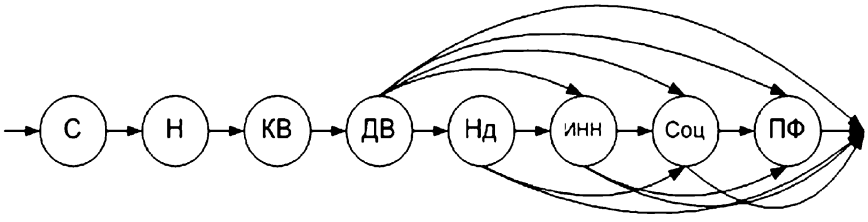


Рис. 11. Схема секции «Реквизиты физического лица»

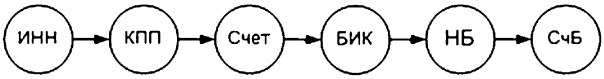


Рис. 12. Схема секции «Реквизиты организации»

аналогичны «Реквизитам физического лица». Разделение типов «Реквизиты организации» и «Реквизиты физического лица» вызвано еще более изменчивой природой (в России эта секция меняется с регулярностью в 1–2 года) и большим отличием составляющих от «Реквизитов физического лица».

Большинство Форм включают в себя базисные секции «Имя» и «Дата», наличие остальных могут варьироваться в зависимости от назначения Формы.

Разделение документы на секции подразумевает то, что существуют методы выделения, идентификации и трансформирования к стандартному виду этих секций в реальном документе. Разработка таких методов и средств представляет из себя отдельную и достаточно сложную задачу, а выделение базисных типов секций предполагает, что либо известны алгоритмы и методики построения, либо такие методы и средства уже построены. Приведем примеры, объясняющие сложность задачи выделения и идентификации для первых трех базисных типов секций: «Имя», «Дата», «Адрес».

1. «Имя» — написание имени, как в различных документах, так и внутри самого документа может, во-первых, серьезно варьироваться, а во-вторых, склоняться. Например, Иванов Петр Данович может быть записан как П. Д. Иванов или как Иванова П. Д.; если в первом случае все очевидно, то во втором случае разобраться, кого имели в виду, Иванова Петра или Иванову Пелагею — можно только проведя контекстный анализ окружения. Кроме этих проблем есть еще проблемы сокращений, например: пишут «А Иванов П в данное время находился ....». Разобраться, кого имели ввиду (А. Иванова или Иванова П.), и правильно выделить имя достаточно сложно.

2. «Дата» — написание даты так же, как и «Имени» достаточно часто подвержено неоднозначности и сокращению. Например, существует несколько форматов написания даты: ДД.ММ.ГГ, или ДД.ММ.ГГГГ, или ДД/ММ/ГГГГ, или ГГГГ-ММ-ДД, или ММ/ДД/ГГГГ и т. д., в этих форматах разняться не только разделители – "./", но и порядок элементов даты. Кроме этого, дату можно записать целиком прописью – «первое июля тысяча девятьсот тринадцатого года» или частично – «1 июля 1913» и т. д. Таким образом, задача выделения даты и приведение к одному виду достаточно сложна.

3. «Адрес» — написание адреса в документах еще более подвержено вариациям и искажениям, более подробно проблема описана выше.

Выводы

Данная статья посвящена построению Формы на основе выделения трех ее основных компонентов: модели содержания, модели взаимодействия и модели визуализации и дальнейшего разложения каждой компоненты на слои обязательные, регламентированные и слои расширения.

Введенное разделение не является общепринятым даже на верхнем уровне. Если понятия аналогичные модели содержания и модели визуализации присутствуют во многих концепциях, то модель взаимодействия присутствует в них в виде набора статических или операциональных описаний конкретных программ ввода/вывода и записи/извлечения данных из базы. Явное выделение модели взаимодействия позволило структурировать необходимые функции, исследовать их с единых позиций и, в ряде случаев, симметризировать. В частности, процесс распознавания – один из основных механизмов ввода бумажных документов, который до сих пор рассматривался как функция – «черный ящик», занял свое место наряду с другими процессами ввода/вывода.

Важным результатом является также выделение обязательных и регламентированных слоев. Оно обеспечивает разработку стандартных интерфейсов, благодаря чему «правильные описания» воспринимаются даже системами их не обрабатывающими. При этом система всегда может быть дополнена (а не переработана) для обработки соответствующих слоев.

Литература

1. *Арлазаров В. В.*, *Постников В. В.*, *Шоломов Д. Л.* Cognitive Forms — система массового ввода структурированных документов // Управление информационными потоками. Сборник трудов Института системного анализа РАН. М.: URSS, 2002.
2. *Емельянов Н. Е.* Виды представления структурированных данных // Теоретические основы информационной технологии / Сб. тр. Вып. 22. М.: ВНИИСИ, 1988.
3. *Богачева А. Е.*, *Емельянов Н. Е.* Семантическая Модель документа // Системные исследования. Ежегодник. М.: URSS. 2003. С. 360–375.
4. *Emelyanov N. E.*, *Solovyev A. V.*, *Schelkacheva I. V.* Classification of Structured Data Representations // Proceedings of the Third International Workshop on Advances in Databases and Information Systems. MEPhI Publishing. 1996. Vol. 2.

5. *Постников В. В.*, Разработка методов наложения формы на графическое изображение документа // Интеллектуальные технологии ввода и обработки информации. М.: URSS, 1998.
6. *Feldbach M.*, *Tönnies K. D.* Word egmentation of Handwritten Dates in Historical Documents by Combining Semantic A-Priori-Knowledge with Local Features // Seventh International Conference on Document Analysis and Recognition Volume I. 2003. P. 333.
7. *Brakensiek A.*, *Rottland J.*, *Rigoll G/* Handwritten Address Recognition with Open Vocabulary Using Character N-Grams // Eighth International Workshop on Frontiers in Handwriting Recognition (IWFHR'02). 2002. P. 357.
8. *Cesarini F.*, *Gori M.*, *Marinai S.*, *Soda G.* A system for data extraction from forms of known class // Third International Conference on Document Analysis and Recognition (Volume 2). 1995. P. 1136.
9. *Cracknell C.*, *Downton A. C.*, *Du L.* An Object-Oriented form Description Language and Approach to Handwritten Form Processing // 4th International Conference Document Analysis and Recognition (ICDAR '97). Volume I and Volume II. 1997. P. 180.
10. *Sholomov D. L.* Syntactical Approach to Post-Processing of Fuzzy recognized Text // Proc. of The International Conference on Machine Learning, Technologies and Applications, June 2003, USA. CSREA Press. P. 115–121.
11. *Sholomov D. L.*, Interpreting the Indistinctly Recognized Textual Constructions // Pattern Recognition and Image Analysis. 2003. Vol. 13. № 2. P. 353–355.
12. *X*Forms 1.0, W3C Recommendation 14 October 2003, [Электронный ресурс], http://www.w3.org/TR/2003/REC-xforms-20031014/

Арлазаров Владимир Викторович. С. н. с. ИСА РАН. К. т. н. Окончил в 1999 г. МИСиС. Количество печатных работ: 6. Область научных интересов: распознавание образов, обработка изображений, системы массового обслуживания. E-mail: vva777@gmail.com

Кривцов Валерий Евгеньевич. С. н. с., декан МФТИ. К. ф.-м. н. Окончил в 1971 г. МФТИ. Количество печатных работ: 94 публикации, 3 монографии. Область научных интересов: информационные технологии и компьютерные науки, математическое программирование, математические модели и методы в экономике. E-mail: krivtsov@phystech.edu

Полевой Дмитрий Валерьевич. С. н. с. ИСА РАН. К. т. н. Окончил в 2004 г. МФТИ. Количество печатных работ: 9. Область научных интересов: методы искусственного интеллекта, оптическое распознавание, системы обработки документов. E-mail: dvpsun@gmail.com

Слугин Дмитрий Геннадьевич. Научный сотрудник ИСА РАН. Окончил в 2000 г. МГУ им. М. В. Ломоносова. Количество печатных работ: 4. Область научных интересов: распознавание образов, обработка изображений, электронный документооборот, распределенные вычисления. E-mail: slugindm@gmail.com

Янишевский Игорь Михайлович. С. н. м. ИСА РАН. К. ф.-м. н. Окончил в 1988 г. МГУ. Количество печатных работ: 16. Область научных интересов: теория случайных процессов, распознавание образов, теория оптимального управления. E-mail: igor_y@cs.isa.ru