

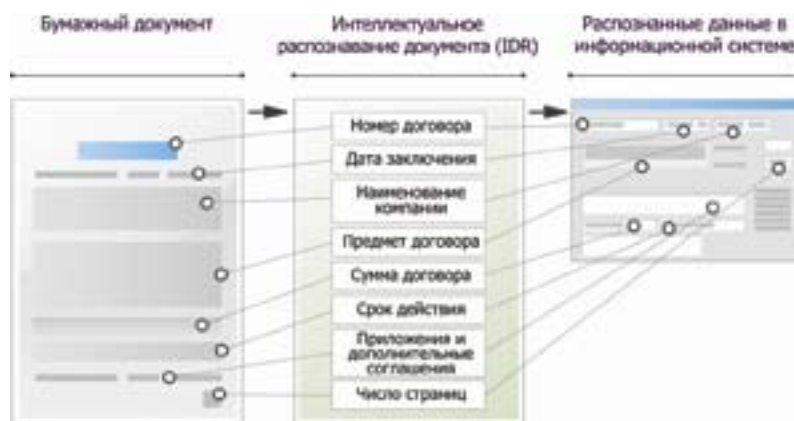
# От распознавания символов — к пониманию документов

По данным шведской фирмы Anoto Group, 86% компаний в мире до сих пор используют для сбора данных бумажные документы, заполняемые вручную. Вместе с тем 75% компаний интересуются решениями, которые могли бы сканировать данные в электронную форму и сразу отправлять их в хранилище. Бизнес нуждается в том, чтобы сведения, содержащиеся на бумажных носителях, стали доступны в электронном виде. Естественно, для этого потребуется перенести их в компьютерные системы, причем задачу переноса поставить не перед человеком, а перед компьютером.

Сегодня уже вряд ли кого удивишь системами оптического распознавания текста и документов — любой человек, даже не обладая выдающейся скоростью печати, может посредством обычного офисного сканера перенести в компьютер в редактируемом виде текст, напечатанный убогим шрифтом на странице формата А4, всего за одну минуту. Но для бизнеса наибольший интерес представляют системы, способные извлекать информацию из деловых документов: договоров, заявлений, анкет, счетов-фактур, платежей, бухгалтерской отчетности и т. п. Такие системы гораздо сложнее обычных «распознавалок», потому что от них требуется распознать не только текст, но и структуру документа. И самое главное — они должны «понимать» содержание текста: например, что вот эти несколько знаков, которые были только что распознаны, означают номер договора, фамилию плательщика или итоговую цену товара и каждое из них нужно поместить в соответствующее поле базы данных информационной системы. Иными словами, решения такого класса реализуют процесс «интеллектуального распознавания документов» (Intelligent Document Recognition, IDR).

Образно говоря, обычные системы распознавания подобны мозгу ребенка, который только видит буквы, но не понимает и не может оперировать содержащейся в тексте информацией. Системы же класса IDR и «системы понимания документов» (Intelligent Document Understanding, IDU) скорее аналогичны мозгу взрослого человека — профессионала в конкретной предметной области, способного выделять в документе различные информационные блоки: даты, финансовые сведения, условия соглашения, сведения об организации, персональные данные и т. д.

Но сложность обработки документа определяется не только совершенством систем ввода, но и типом самого документа. В зависимости от структуры, документы можно разделить на следующие типы:



- структурированные: анкеты, страховые полисы, декларации и т. п.;
- слабоструктурированные: накладные, счета-фактуры, платежные поручения и т. п.;
- неструктурированные: договора, письма, статьи и т. п.

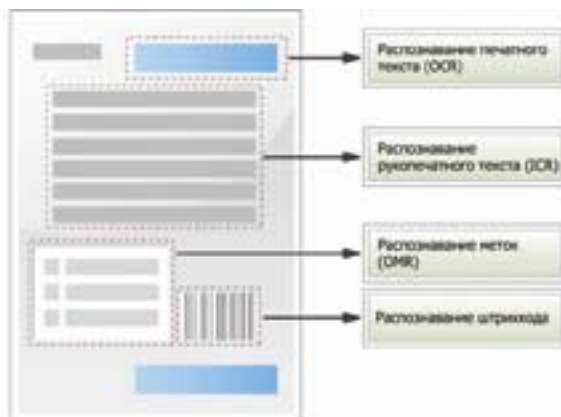
Для каждого типа документа применяются свои методы поиска областей, в которых содержатся данные.

## Структурированные и слабоструктурированные документы

Проще всего обстоит дело со структурированными документами, так как для них заранее определены поля с данными, размеры этих полей и их взаимное расположение. Системе достаточно совместить изображение документа с его геометрическим шаблоном, чтобы понять, из какого участка текста требуется распознать данные. Решения этого типа отно-

сятся к системам оптического распознавания документов (Optical Document Recognition, ODR).

Слабоструктурированные документы обработать сложнее: состав и взаимное расположение информационных полей и блоков здесь определены заранее, а вот их размеры и точное расположение неизвестны. И в этом случае наложением шаблона не обойтись, требуется учитывать соответствие распознанного текста синтаксису реквизита, его формату, наличие рядом ключевых слов. Так, для поля «Назначение платежа» платежного поручения характерным моментом является наличие в нем аббревиатуры «НДС», например: «в том числе НДС 18% — 1140-00» либо «НДС не облагается». То есть в этом случае речь идет уже о синтаксических методах распознавания документов. Такого рода технологии относятся к классу IDR.



Распознавание структурированного документа происходит с помощью наложения геометрического шаблона

Примером применения технологий IDR может служить автоматизация ввода заявлений на получение биометрического загранпаспорта в государственной автоматизированной системе паспортно-визовых документов, которая успешно работает в отделениях ФМС России. Не каждому, наверное, известно, что расположение и размер полей заявлений на выдачу загранпаспорта не унифицированы и могут меняться в зависимости от способа печати. На бланках, распечатанных в типографии, расположение одно, у скачанных с сайта ФМС — другое, а у бланка, взятого с какого-либо другого сайта или изготовленного самостоятельно в любом удобном текстовом редакторе, — третье.

Имеют значение и некоторые структурные особенности: например, один документ или целая таблица с данными могут располагаться на нескольких листах бумаги. Естественно ожидать от систем класса IDR корректной обработки и в этих случаях.

## Неструктурированные документы — будущее систем IDU

С точки зрения обработки неструктурированные документы представляют собой наибольшую сложность. Реквизиты, которые необходимо из них извлечь, распределены в произвольном порядке и могут быть оформлены в виде отдельных полей или располагаться в тексте. К неструктурированным документам можно отнести и многостраничные документы, в которых количество страниц не детерминировано.

Основная проблема сегодня состоит в описании структуры таких документов. Системы, способные их обработать, должны обладать мощным искусственным интеллектом и быть способны к самообучению. Ведущие разработчики систем распознавания и понимания документов лишь сравнительно недавно начали получать положительные результаты в решении задачи создания таких систем.

## Расознавание и понимание документов: мировой и российский рынки

Мировой рынок решений распознавания и понимания документов (на Западе этот класс систем называют document capture and form processing), по разным оценкам экспертов, составляет сегодня \$1–1,5 млрд в год. Основными игроками на нем выступают американские и европейские ком-

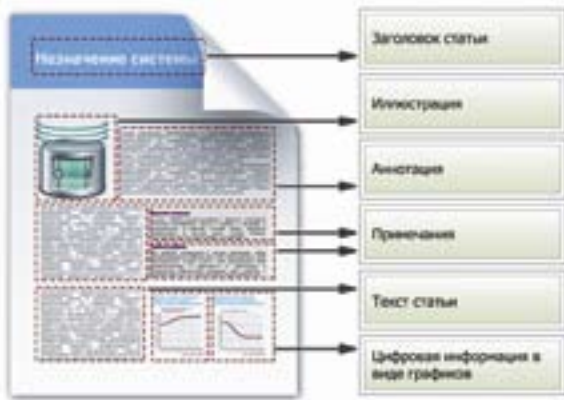


Пример «плавающего» поля — «номер договора»

пании — Kofax, Cardiff, ReadSoft, EMC, AnyDoc Software и др. Российский же рынок, по разным оценкам, составляет порядка \$10–15 млн в год, и на нем лишь два основных игрока — отечественные компании Cognitive Technologies и АВВУ, а западные компании (в основном Kofax и EMC) мало заметны. Потенциал отечественного рынка значителен: динамика его роста, по прогнозам, способна составить от 20 до 30% уже в ближайшие два-три года.

Компания АВВУ автоматизировала ввод налоговых деклараций для ФНС России. Ввод анкет московских школьников, студентов и пенсионеров для выдачи льготных проездных билетов в Московском метрополитене по сей день осуществляется с помощью системы Cognitive Forms, а результаты ЕГЭ обрабатываются системой АВВУ.

Одно из основных отличий отечественного рынка распознавания и понимания документов от западного состоит в том, что на Западе



Пример неструктурированного документа — журнальная статья

Стоит отметить, что российские программные продукты по своим функциональным характеристикам явно не хуже западных аналогов. Низкое качество заполнения документов (на Западе культура письма печатными буквами прививается с детства), слабый уровень печати документов (случается, что линии разграфки наезжают на полезную текстовую информацию), а также более высокая сложность русского языка с точки зрения его компьютерной обработки (например, в европейских языках отсутствуют двухкомпонентные буквы, как буква «ь») вынуждают разработчиков изобретать методы и алгоритмы с гораздо более высоким уровнем искусственного интеллекта.

Первые решения по распознаванию документов стали появляться у нас в стране уже более 15 лет назад. В 1994 году Cognitive Technologies разработала систему ввода налоговых деклараций и справок о доходах для РГНИ Республики Башкортостан, а с 1998 года в Пенсионном фонде России работает система ввода анкет застрахованного лица.

весьма активно внедряются безбумажные решения (электронная бумага, планшетные компьютеры) и более распространены решения по обработке электронных форм документов, заполняемых непосредственно через Интернет.

Стоит отметить, что решение задач понимания неструктурированной информации в мире ведется параллельно как разработчиками систем автоматизации документооборота, так и разработчиками поисковых систем, главным образом «Яндекс» и Google. Известно, например, что компания «Яндекс» приобрела морфологические разработки у компании АВВУ и синтаксический анализатор текста у Cognitive Technologies. По мнению одного из основателей Cognitive Technologies, члена-корреспондента РАН В.Л. Арлазарова, первые оптимальные результаты в области создания систем обработки неструктурированной информации мы сможем увидеть уже в ближайшие пять лет.

Алексей Бодров,  
Cognitive Technologies

Fujitsu Primergy TX100 S2 Core Edition

Компания Fujitsu анонсировала выпуск Primergy TX100 S2 Core Edition — удобного в эксплуатации и недорогого сервера стандартной архитектуры, оптимизированного для компаний малого бизнеса. Специальная технология, обеспечивающая доступ к серверу, который не имеет собственного монитора, мыши и клавиатуры, помогает экономить пространство. Функция Just power-up'n'run («включи и работай») автоматически устанавливает серверное ПО, а администрирование может выполняться с любого ПК, подключенного к сети. В сервере предусмотрены функции самовосстановления, предустановлена ОС Microsoft Windows Server 2008 R2 Foundation. Конфигурация новинки включает в себя процессор Intel Xeon X3430, два жестких диска по 1 Тбайт каждый, 4 Гбайт оперативной памяти DDR3 и встроенный привод DVD.



Getac V200

Компания Getac объявила о снятии с производства защищенного планшетного ноутбука V100–2M и замене его моделью V200, базирующейся на передовой архитектуре Intel Calpella. Новинка оснащена процессором Intel Core i7–620LM с частотой 2 ГГц и возможностью повышения частоты до 2,8 ГГц благодаря технологии Intel Turbo Boost. В отличие от предыдущей модели, видеоконтроллер новинки встроен в процессор, а не в чипсет. Getac V200 может быть оснащен оперативной памятью объемом до 8 Гбайт и твердотельным диском объемом 80 Гбайт. Возможность работы при отрицательных температурах включена в базовую конфигурацию устройства. Как и в предыдущей модели, обеспечивается поддержка фирменной технологии Getac QuadraClear с яркостью экрана в 1200 нт и сенсорного экрана multi-touch, совмещенного с дигитайзером (работать с которым можно даже в перчатках). Новинка соответствует стандартам MIL-STD-810G и IP65 для защиты от грязи, пыли, воды, вибрации, перепадов температуры и других вредных факторов.



Конференция «ИТ для предприятий металлургии»

Темы конференции:

- Стратегические задачи для СЭО в условиях восстановления экономики
- Повышение эффективности управления и сокращение производственных издержек
- Определение и использование ключевых показателей эффективности (KPI)
- Информационные панели руководителей, основные показатели металлургического производства
- Оптимизация инфраструктуры и управление ИТ-услугами
- Формирование цепей поставок, утилизация активов, управление взаимоотношениями с клиентами
- Изменение роли ИТ-директора современного предприятия

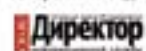
Москва  
отель «Президент»  
11 ноября

Дополнительная информация и регистрация на сайте <http://www.idc-cema.com/events/metal10> и по телефону +7 495 661 6166.

Золотые партнеры:



Генеральный медиа партнер:



Информационная поддержка:



реклама