УДК 004.418

АВТОМАТИЗАЦИЯ РАСПОЗНАВАНИЯ ДОКУМЕНТОВ И ОБРАБОТКИ РАСПОЗНАННЫХ ДАННЫХ НА ОСНОВЕ РЕШЕНИЯ KOFAX TRANSFORMATION MODULES

Окольнов Юрий Викторович[1], Поддубный Михаил Александрович[2], Тятюшкина Ольга Юрьевна[3]

[1]Программист;
ООО «Логстрим»;
Россия, 141981, Московская обл., г. Дубна, пр-т Боголюбова, д. 26, оф. 4;
e-mail: yokolnov@logstream.ru.

[2]Магистрант;
ГБОУ ВО МО «Университет «Дубна»,
Институт системного анализа и управления;
Россия, 141980, Московская обл., г. Дубна, ул. Университетская, д. 19;
e-mail: mpoddubniy@logstream.ru.

[3]Кандидат технических наук, доцент Института системного анализа и управления;
ГБОУ ВО МО «Университет «Дубна»,
Институт системного анализа и управления;
Россия, 141980, Московская обл., г. Дубна, ул. Университетская, д. 19;
e-mail: tyatyushkina@mail.ru.

В данной статье описана технология Kofax Transformation Modules компании Kofax, которая позволяет автоматизировать и ускорить распознавание и проверку деловой документации при ее потоковом сканировании.

Ключевые слова: Kofax Transformation Modules, распознавание, автоматизация обработки, деловая документация.

AUTOMATISATION OF DOCUMENT RECOGNITION AND PROCESSING OF RECOGNISED DATA BASED SOLUTIONS KOFAX TRANSFORMATION MODULES

Okolnov Yurij[1], Poddubnyi Mihail[2], Tyatyushkina Olga[3]

[1]Programmer;
Ltd «Logstream»;
Russia, 141981, Moscow reg., Dubna, office 4 26 ave. Bogolubova;
e-mail: yokolnov@logstream.ru.

[2]Graduate student of the Department of System Analysis and Management;
Dubna State University,
Institute of the system analysis and management;
Russia, 141980, Moscow reg., Dubna, 19 University st.;
e-mail: mpoddubniy@logstream.ru.

[3]Candidate of Science in Engineering, Associate professor of the Department of System Analysis and Management;
Dubna State University,
Institute of the system analysis and management;
Russia, 141980, Moscow reg., Dubna, 19 University st.;
e-mail: tyatyushkina@mail.ru.

This article describes techonological system Kofax Transformation Modules produced by Kofax. That system performs streaming scanning of business documentation and automatizes its recognition and processing while also improving speed of manual operations.

Keywords: Kofax Transformation Modules, recognition, automatic processing, business documentation.

Введение

Современные компании часто сталкиваются с проблемами обработки большого количества документации, при этом одни и те же операции с документами осуществляются по несколько раз: сканирование, регистрация, проверки. Как следствие: обработка документов требует существенных затрат (денежных и временных); качество документов и качество данных в информационных системах (ИС) компаний существенно зависит от влияния человеческого фактора; данные в ИС компании дублируются.

На рынке услуг по автоматизации оптического распознавания и обработки документов выступают такие компании как Abbyy, Kofax, Captiva. Не все компании обладают таким богатым инструментарием как компания Kofax. Ниже представлен один из наиболее востребованных продуктов Kofax.

1. Описание продукта Kofax Transformation Modules

Kofax Transformation Modules (KTM) – наиболее полное и универсальное из представленных на современном рынке решений классификации документов и извлечения из них данных. Оно обеспечивает бесконтактную обработку (по технологии Touchless ProcessingTM) печатных и рукописных форм, счетов-фактур, чеков, договоров, корреспонденции и документов любого другого типа на одной платформе, предлагая своим пользователям во всем мире не имеющие аналогов преимущества для бизнеса. KTM предоставляет уникальные технологии классификации, разделения, извлечения данных и их проверки с поддержкой обучения системы. Эти технологии обеспечивают точную автоматическую обработку даже самых сложных документов [1].

Разделение, извлечение и проверка данных вручную – медленный, дорогостоящий и подверженный ошибкам процесс. KTM эффективно решает эти задачи за счет следующих преимуществ:

Обеспечение решений с поддержкой бесконтактной обработки для ускорения определения дальнейших пунктов назначения информации

Автоматизация классификации и разделения документов, извлечения и проверки информации.

Ускорение обработки, сокращение затрат, повышение качества данных.

Платформа KTM предназначена для крупных и средних предприятий, использующих медленные, дорогостоящие и подверженные ошибкам ручные процессы классификации и разделения документов различных типов, извлечения и проверки многочисленных элементов данных из этих документов (как печатных, так и электронных) для передачи в бизнес-процессы в нисходящем направлении. KTM поможет предприятиям:

- Выполнить ввод и обработку любого документа – KTM поддерживает документы и информацию любого типа, обеспечивая единое решение для ввода данных в среде всей организации, что способствует повышению рентабельности. Многие из предлагаемых конкурентами решений предназначены для использования в конкретных подразделениях и работают лишь с определенными типами документов.
- Реализовать бесконтактную обработку в максимальном объеме – KTM использует передовые технологии распознавания, обеспечивающие беспрецедентную скорость извлечения данных из текста любого типа (печатного, рукописного и скорописного) на более чем 140 языках. Высокая скорость извлечения данных позволяет уменьшить количество выполняемых вручную операций, сократить количество выполняющих эти операции персонала и более эффективно использовать возможности бесконтактной обработки.

- Свести к минимуму затраты на предварительную настройку – KTM использует при классификации, разделении и извлечении технологии самообучения на примерах, что позволяет оперативно подготовить систему к работе с документами любых типов. Многие из предлагаемых конкурентами решений используют сложные правила и шаблоны документов, подготовка которых занимает несколько недель или даже месяцев. Другие решения используют автоматические шаблоны, не поддерживающие работу с письмами, отчетами и другими неструктурированными документами.
- Свести к минимуму объем текущих затрат на обслуживание – KTM продолжает процесс самообучения в процессе обработки документов, оперативно увеличивая скорость классификации и извлечения информации. Это способствует увеличению рентабельности и устраняет необходимость в ручном конфигурировании системы в случае изменения документов или ввода новых документов.
- Оптимизировать обработку ошибок – KTM предлагает современные интуитивно понятные пользовательские интерфейсы с возможностью настройки. Они позволяют быстро и эффективно устранять возникшие ошибки [1].

2. Настройка распознавания и обработки документов

KTM Project Builder это интегрированная среда для настройки и разработки пользовательского функционала системы KTM с графическим интерфейсом.

Project Builder позволяет настроить функционал как используя готовые методы, так и создать свои подходы, программируя на языке WinWrap Basic. Также можно подключать библиотеки, написанные на других языках через механизм COM-объектов.

Project Builder позволяет гибко настраивать полнотекстовое распознавание (ПТР), классификацию, разделение документов, извлечение данных из распознанного текста, проверку данных различными способами.

Для полнотекстового распознавания в KTM есть несколько библиотек (движков) сторонних разработчиков. Для русскоязычного распознавания могут использоваться две из них – это Abbyy FineReader и Recostar.

Оба движка имеют широкие настройки: выбор языка распознавания, ограничение набора распознаваемых символов до только цифр, или только букв. Также можно настроить игнорирование типичных нетекстовых элементов отсканированных изображений – узоры, линии, точки от преобразования цветов в черно-белое изображение.

Как было отмечено выше, актуальными являются задачи по автоматизации распознавания и последующей обработке документов. С помощью KTM они решаются в следующей последовательности.

Классификация документов может производиться как с помощью «обучения» проекта, так и путем настройки методов и правил разработчиком.

Для классификации в проекте заводится иерархический набор классов, отражающий значимые для проекта типы документов.

Обучаемая классификация позволяет очень просто создать набор «образцов» для каждого класса. С этими образцами система будет сравнивать документы, приходящие на обработку в систему. При этом можно выбрать как сравнение «визуальное», по графическому шаблону, так и сравнение по содержимому, по тексту документа.

Также в процессе работы пользователей с системой они могут нажатием одной кнопки отправлять неклассифицированные образцы для дальнейшего «обучения» системы.

Вместо классификации по содержимому можно использовать классификацию по заданным шаблонным фразам, найденным или не найденным на документе. Каждой из фраз можно задать вес.

В случае специфических требований можно не использовать типовую классификацию, а назначать классы программным кодом, базируясь на найденных на документе штрих-кодах, или на расположении страниц в документах или любых других формализованных подходах.

Каждый класс в проекте имеет набор заданных полей. Значения полей задаются так называемыми локаторами. Локаторы – это готовые методы для извлечения данных с документа. Перечислим несколько самых часто используемых.

Локатор формата. Этот локатор ищет в ПТР заданные регулярные выражения в заданной части документа с заданными ключевыми словами вокруг.

Локатор зон. Этот локатор нацеливается на конкретную позицию документа и в небольшой зоне производит более точное распознавание текста. При этом можно включить различные фильтры изображения для удаления точек, линий, утолщения или утоньшения символов. Можно выбрать один из движков распознавания (FineReader или Recostar).

Можно выбрать способы изображения символов – машинная печать или буквы, нарисованные от руки. Также частично поддерживается распознавание ручного написания букв. Но оно не является надежным в связи с высокой вариативностью способов написания букв (почерков). При этом ручное написание цифр распознается достаточно хорошо.

Локатор групп. Этот локатор является обучаемым и его настройка сводится к тому, что специалист явно указывает проекту, где искать данные, выделяя их мышкой. Путем перебора нескольких десятков документов создается база «обучения». Когда проект сдается в работу обучение может быть продолжено уже пользователями, которые аналогичным образом выделяют данные на документе, если система не смогла их обнаружить. После этого документы автоматически добавляются в базу обучения.

Локатор базы данных. Данный локатор используется, когда на документе нужно найти сгруппированные данные, соответствующие какой-то записи в базе данных, например ФИО человека и другие его параметры. При этом возможно нечеткое сравнение строк. Для этого должен быть установлен сервер сверки, который индексирует заданные таблицы в базе данных и проводит быстрое нечеткое сравнение.

После извлечения данные могут быть проверены правилами форматирования и валидации.

Форматирование использует готовые методы, а также скрипты, чтобы проверить формат данных, поправить данные, выдать пользователю предупреждение. Например, можно проверять формат дат, числовых полей, проверять данные, где есть «сумма сверки» (checksum)

Правила валидации позволяют настроить взаимопроверку нескольких полей и выдать соответствующие предупреждения. Например, можно взаимно сверить суммы в финансовых документах, которые являются пропорциональными друг другу – сумму НДС, сумму без НДС, сумму с НДС.

3. Пример внедрения продукта

Один из крупных проектов по внедрению технологии KTM был реализован для российского оператора сотовой связи по обработке первичных финансовых документов таких как акты, счета-фактуры, договора, товарные накладные, универсальные передаточные документы и другие. В ходе проведения работ по извлечению атрибутов полей с помощью KTM Builder были реализованы разнообразные функции и методы системы.

Целью данного проекта являлось сравнение версий документа, отправленного контрагенту и присланного обратно. Для этого извлекались наиболее важные данные с возвращенного документа и сравнивались с данными в БД заказчика. При этом для каждого из видов документов имеется своя специфика.

Счета-фактуры являются достаточно стандартизированным документом. Верхняя, шапочная часть обычно хорошо распознается локаторами формата, привязанными к опорным словам полей.

Также часть данных распознается локаторами зон, которые подстраиваются скриптовым кодом для уточнения положения зон точного распознавания относительно опорных слов.

Для нижней табличной части приходится применять созданные нами библиотеки, которые анализируют и извлекают табличные данные. Наличие подписей внизу документа проверяется локаторами зон в режиме «галочка», который проверяет заполненность знакоместа пикселями.

Акты, напротив, являются почти не стандартизированным документом и каждый из множества контрагентов имеет несколько видов специфических актов, расположение данных на которых сложно предугадать. При этом и опорные слова могут быть весьма разнообразными, и формат данных не всегда одинаков. Для актов настраиваются гибкие варианты распознавания с попыткой предусмотреть большую часть вариантов. Для распознавания используются локаторы формата, которые ищут необходимые данные по различным регулярным выражениям, а также групповые локаторы, которые пытаются обнаружить сгруппированные по положению пропорциональные числа, являющиеся итоговыми суммами на документе.

Договора в этом проекте не проходят извлечение отдельных данных. Для договоров берется весь ПТР отосланного документа и сравнивается с ПТР возвращенного документа. На выходе создается протокол, где различия в тексте помечены цветом.

Заключение

В процессе реализации проекта была доказана эффективность применения КТМ для обработки финансовых документов.

Помимо применения технологии КТМ в обработке финансовых документов ее можно использовать для автоматизации обработки документов страховых компаний, банковских организаций, муниципальных учреждений и прочих организаций.

Список литературы

1. Официальный сайт компании Kofax® Limited (NASDAQ: KFX). – [Электронный ресурс]. URL: http://www.kofax.com.ru/software/transformation/features.php (дата обращения: 13.03.2016).
2. Гуржеева Н., Пушкарук Д. Потоковый ввод и распознавание с Kofax. – [Электронный ресурс]. URL: http://www.slideshare.net/oksanakurysheva/kofax-alfresco-moscow-meetup (дата обращения: 13.03.2016).