Метод классификации распознанных страниц деловых документов на основе текстовых ключевых точек*

О.А. Славин[I,II], В.Л. Арлазаров[I,II]

[I] Институт системного анализа Федерального исследовательского центра «Информатика и управление» Российской академии наук, Москва, Россия
[II] ООО «Смарт Энджинс Сервис», Москва, Россия

Аннотация. В работе рассматривается задача классификации распознанных страниц деловых документов. Деловые документы, используемые в документообороте, в том числе в обмене документами между организациями, обладают определенной стандартизацией, они могут быть как неструктурированными, так и структурированными. В банках или страховых компаниях часто необходимы такие документы как доверенность, договор, карточка с образцами подписей и печатей, устав, контракт, счет, свидетельства о регистрации и т.п. При создании и ведении электронных архивов бумажные документы оцифровываются, а цифровые образы страниц (сканы страниц) могут быть распознаны и проанализированы. Одной из задач анализа является классификация образа страницы, состоящая в проверке его принадлежности определенному классу. Предложен простой метод классификации деловых документов, дающий приемлемые результаты.

Ключевые слова: *классификация текстов; распознавание документов; OCR; ошибка распознавания; template matching.*

DOI: 10.14357/20790279180504

Введение

Важнейшей функцией современных систем ввода документов в компьютер является распознавание структуры и содержимого оцифрованных документов, а также анализ текста, извлеченного из печатного документа. Эти задачи может быть существенно успешнее решены, если структура документа известна заранее.

Будем называть деловыми (административными документами [4]) документы, участвующие в некотором делопроизводстве, предназначенные для хранения в электронных архивах. Совокупность атрибутов делового документа является карточкой документа, представляющей интерес при поиске документа и анализе его содержимого. Примерами деловых документов являются: свидетельства о регистрации, документы для удостоверения личности (далее – идентификационные документы), доверенности, счета, заявления, примеры фрагментов которых приведены на рис. 1. От деловых будем отличать произвольные документы, не содержащие атрибутов, необходимых в каком-то делопроизводстве, и представляющих интерес как приложения к деловым документам.

По способу изготовления документы подразделяются на полиграфические и напечатанные. Полиграфические документы, такие как паспорт гражданина РФ, дипломы ВУЗов, права, свидетельства ЕГРЮЛ/ЕГРИП, напечатаны на специальной бумаге, обладающей элементами защиты от подделок (защитные волокна, водяные знаки). Полиграфические документы являются стандартными или стандартизованными и изготавливаются специализированными организациями. Множество типов документов, таких как справки, доверенности, приказы, заявления, печатаются на лазерных принтерах гражданами или сотрудниками организаций. В зависимости от способа заполнения документы могут быть рукописными или печатаными, однако рассматриваемые в работе рукописные документы (например, анкеты) используют бланк с печатными статическими текстами.

Документы могут быть одностраничными и многостраничными.

Документ будем рассматривать в виде двух слоев: шаблон документа и заполнение документа. Заполнение содержит уникальную информацию документа в виде однострочных или многострочных полей. Шаблон документа одинаков для всех экземпляров, состоит из статических элементов: линий, таблиц, статических текстов, изображений, полей пометок (чек-боксов), маркеров, сложного фона.

* Работа выполнена при частичной финансовой поддержке грантов РФФИ (проекты № 16-07-00616 и 16-07-01051).

Рис. 1. Примеры деловых документов (доверенность, счет, свидетельство о регистрации, заявление)

Шаблоном документа может называться готовый лист для печати на нем заполнения или шаблон в электронной форме (например, в формате PDF).

Будем различать жесткоструктурированные и сложноструктурированные документы. Под жесткоструктурированными понимаются документы, шаблон которых создается таким образом, чтобы при печати сохранялись точные расстояния между любыми элементами разметки для любых экземпляров напечатанных документов (с точностью до допустимого коэффициента масштабирования). при заполнении содержится в специально отведенных для этого зонах. Примером жесткоструктурированного документа является анкета-заявление на выдачу загранпаспорта РФ. В сложноструктурированных документах статические элементы и элементы заполнения не привязаны к определенному месту и не могут иметь близкие геометрические характеристики. Примером сложноструктурированного документа является счет (invoice), атрибуты которого могут размещаться по разному по отношению к другим атрибутам. Деловые документы могут быть как жесткоструктурированными, так и сложноструктурированными. Однако многие классы деловых документов имеют простую структуру, в которой немногочисленные заголовки и ключевые слова применены при изготовлении документа намеренно для лучшего его узнавания человеком.

Постановка задачи классификации документа (или образа документа) состоит в следующем. Пусть имеется несколько моделей $\{M_1, M_2, \dots M_n\}$, каждая из которых соответствует некоторому классу документов, и известна функция расстояния между документом D и моделью M: $r(D, M)$. Необходимо ранжировать модели $\{M_1, M_2, \dots M_n\}$ по расстоянию между документом D и каждой из моделей M_i, результат представляется оценками всех моделей:

$$r_1=r(D, M_{i_1}), r_2=r(D, M_{i_2}), \dots r_n= \\ =r(D, M_{i_n}), 1\le i_p\le n, r_1\le r_2\le \dots \le r_n. \quad (1)$$

После ранжирования рассматриваются только модели, расстояние до которых не превышает некоторого порога m:

$$r_{1=}r(D, M_{i_1}), \dots r_{q=}r(D, M_{i_q}), \\ 1\le i_p\le n, r_1\le \dots \le r_q\le m, 1\le q\le n, \quad (2)$$

либо выбирается ближайшая к документу модель M_{i_1}.

В работе будет рассмотрен способ построения модели и выбора функции расстояния между документом и моделью для распознанных документов, в которых имеются ошибки распознавания. Существенным для описываемого способа является использование информации о координатах распознанных слов.

1. Обзор существующих методов классификации документов

В работе [1] методы классификации документов подразделяются на методы анализа их структуры, методы анализа текстовой информации и анализа изображений. Анализ структуры документов ориентирован на такие жесткострук-

турированные документы как анкеты или счета-фактуры, основан на поиске статических графических элементов (разделяющие линии, рамки) и статические тексты.

Авторы работы применяют методы анализа изображений для распознавания идентификационных документы, поскольку каждый класс идентификационных документов обычно имеет характерную графическую структуру в виде неизменных текстовых зон (метки полей: имя, фамилия и т. д.), полей с переменным текстом (персональные данные) и одного и того же фона. Для каждого класса идентификационных документов создается одна модель. При создании модели вначале определялись границы полей. Затем вне границ полей извлекаются особые точки (ключевые точки) [3], наличие которых вероятно на всех образцах документа. Особые точки характеризовались с помощью дескрипторов (SIFT, SURF, ORB), то есть локальных описаний, позволяющих отделить особые точки от соседних точек изображения. Некоторые из дескрипторов обладают свойством инвариантности к аффинным преобразованиям, таким как смещение, масштабирование, вращение. При работе с обучающим набором сохраняются только особые точки, которые соответствуют каждому обучающему изображению одного класса. Авторы работы [1] указывают преимущество схемы такого обучения: новая модель документа добавляется независимо от уже созданных моделей. *i*-я модель обозначается множеством M_i из n_i ключевых точек, где $M_i = \{D_{i,1}, D_{i,2}, ..., D_{i,ni}\}$ и D_{ij} – это набор извлеченных дескрипторов для ключевой точки j.

Классификация изображения выполняется на основе описания его особых точек. Он состоит в поиске класса-победителя, который имеет максимальное количество особых точек, наилучшим образом соответствующих ключевым точкам запроса.

Особые точки изображения сопоставляются со всеми изученными моделями, то есть все модели конкурируют друг с другом в этой фазе соответствия. Набор прямого соответствия с моделью M_i обозначается m_i. Модели ранжируются по соотношению $r_i = |m_i|/|M_i|$, а модель с самым высоким коэффициентом $\text{argmax}_i(r_i)$ будет классом-победителем. Далее для всех моделей проводится обратное сопоставление особых точке моделей с особыми точками изображения-запроса. При этом применяется алгоритм RANSAC [2] для поиска геометрического преобразования, которое отображает наибольшее количество пар особых точек и исключает выброс.

Авторы работы [1] сообщают о том, что не существует общедоступных тестовых наборов данных идентификационных документов, и описывают три собственных набора данных для экспериментов, в которых показывается точность классификации – примерно 95%.

Анализ текстовой информации базируется на глобальных дескрипторах текстового контента (например, таких как «мешок слов» или «векторное представление слов» [7]), который затем анализируется классическими классификаторами. В работе [4] предложена архитектура для классификации страниц в банковском процессе. Используются как графические, так текстовые дескрипторы. Графические дескрипторы вычислялись на основе распределения интенсивностей пикселей. Текстовые дескрипторы формировались на основе латентно-семантического анализа для представления содержимого документа в виде сочетания тем. Авторами были оценены несколько готовых классификаторов и различные стратегии для объединения графических и текстовых представлений. Предлагаемый метод был протестирован на наборе реальных деловых документов объемом в 70 000 страниц. Наилучшая точность, полученная при комбинировании нескольких методов классификации, составила 95,6%.

Эффективными методами анализа текстовой информации является применение вероятностных тематических моделей, призванных определить тематику коллекции документов, представляя каждую тему дискретным распределением вероятностей слов, а каждый документ – дискретным распределением вероятностей тем [5, 6]. При анализе документа тематическое моделирование разделяет документ между несколькими темами аналогично (2). Неявно предполагается, что в документе содержится достаточное число слов для построения дискретным распределением вероятностей слов.

В работах [7, 8] описывается аддитивная регуляризация тематических моделей (BigARTM), то есть многокритериальная оптимизация взвешенной суммы критериев, что необходимо для доопределения и обеспечения устойчивости задачи построения тематической модели по коллекции документов.

Перед построением тематических моделей документов на естественном языке обычно проводится нормализация [8] с помощью нескольких преобразований:

- лемматизация – приведение каждого слова в документе к его нормальной форме,
- стеммипг – отбрасывание окончаний и других изменяемых частей слов,
- удаление стоп-слов,
- удаление редких слов и строк,

− выделение ключевых фраз (словосочетаний, характерных для предметной области),
− распознавание именованных сущностей.

Перечисленные преобразования могут быть проведены в автоматическом виде и с применением методов машинного обучения. Нормализованные документы представляются последовательностями термов, в их качестве могут выступать слова, нормальные формы слов, словосочетания.

Несколько предположений:
− о существовании тем,
− о возможности представления документ мультимножеством слов («мешок слов», BoW),
− о вероятностном порождении данных (независимом порождении друг от друга, троек: документ – слово – тема),
− об условной независимости (появление слов в документе по теме зависит от темы, но не зависит от документа, и описывается общим для всех документов распределением)

позволяют формализовать процесс порождения коллекции документов по известному распределению и задачу тематического моделирования [8].

Представление документа в виде «мешка слов» может быть заменена на более адекватные модели, учитывающие порядок слов, такие как статистически устойчивые *n*-граммы и векторные представления слов. Для предварительного сокращения словарей *n*-грамм могут быть использованы методы поиска коллокаций TopMine [9] или SegPhrase [10]. Простой метод TopMine линейно масштабируется и позволяет формировать словарь, в котором каждая *n*-грамма обладает тремя свойствами:
– имеет высокую частоту в коллекции;
– состоит из слов, неслучайно часто образующих *n*-грамму;
– не содержится в (*n*+1)-граммах, обладающих двумя приведенными выше свойствами.

Полезным подходом тематического моделирования коротких текстов считается *тематическая модель битермов* [11]. *Битермом* называется пара слов, встречающихся близко – в одном коротком сообщении или в одном предложении, или в окне *из нескольких* слов. В отличие от биграммы, между двумя словами битерма могут находиться другие слова. Конкретизация «близкого расположения» зависит от постановки задачи и особенностей коллекции документов.

Подходы на основе тематического моделирования могут быть использованы для решения задач классификации деловых документов. Однако описанные подходы к классификации деловых документов нуждаются в совершенствовании по следующим причинам:
– большое число термов в деловых документах, соответствующих нескольким темам, не позволяют достигнуть приемлемой точности классификации;
– наличие ошибок распознавания, усложняющих нормализацию текста.

2. Постановка задачи классификации распознанных деловых документов

Простейшая постановка задачи классификации деловых документов (1) и (2) предполагает, что документ является одностраничным. Обобщением является постановка распознанных многостраничных деловых документов. Пусть имеется последовательность распознанных страниц нескольких документов

$$P=\{P_1, P_2, \ldots, P_m\},$$

и страницы каждого документа не перемежаются со страницами других документов. Последовательность P необходимо разбить на несколько подпоследовательностей, соответствующих отдельным документам

$$D_1=\{P_{D_1,1}, \ldots, P_{D_1,l(D_1)}\}, D_2=\{P_{D_2,1}, \ldots, P_{D_2,l(D_2)}\}, \ldots,$$
$$D_k=\{P_{D_k,1}, \ldots, P_{D_k,l(D_k)}\}$$

причем некоторые документы D_j могут быть одностраничными: $l(D_j)=1$. Для каждого документа D_j необходимо определить расстояние до нескольких имеющихся моделей $\{M_1, M_2, \ldots M_n\}$ и отранжировать модели:

$$r_1=r(D_k, M_{i_1}), r_2=r(D_k, M_{i_2}), \ldots r_n=r(D_k, M_{i_n}),$$
$$1\le i_p\le n,\ r_1\le r_2\le \ldots \le r_n. \quad (3)$$

В данной работе мы будем рассматривать сложноструктурированные документы, которые невозможно классифицировать набором особых точек алгоритмами, аналогичными рассмотренными выше [1]. Отметим, что многие деловые документов относятся к коротким текстам и могут быть классифицированы механизмами битермов и *n*-грамм, однако при этом необходимо учитывать возможные ошибки распознавания.

Для анализа текстовой информации, содержащейся в образе таких документов можно использовать программы распознавания произвольного текста (OCR). В настоящее время одним из самых популярных решений является OCR Tesseract, например, в диссертации [12]. OCR была выбрана для распознавания корпуса архивных документов по следующим причинам: возможность свободного распространения, а также представление результатов распознавания в формате HOCR (HTML OCR) с сохране-

нием информации о координатах распознанных слов.

OCR Tesseract в зависимости от степени зашумления текста и собственных способностей может распознавать образы страниц успешно или неуспешно. Характерные ошибки OCR Tesseract можно разделить на несколько видов:

- E_1 – полный отказ от распознавания страницы, вплоть до того, что не распознано ни одного слова;
- E_2 – число ошибок столь велико, что человек не может понять смысл документа по распознанному текстовому представлению;
- E_3 – неверно распознана структура страницы, например, когда расположенные рядом фрагменты текста в результате оказались разнесенными на значительное расстояние;
- E_4 – наличие небольшого числа ошибок, когда в большинстве неверно распознанных слов присутствует 1-2 ошибки.

На рис. 2 приведен пример зашумленного изображения и результатов распознавания с ошибками. В этой работе мы будем рассматривать классификацию страниц только таких документов, которые массово распознаются OCR Tesseract, в предположении о вероятном возникновении ошибок типа E_3 и E_4.

3. Описание моделей и сопоставления с моделями

Мы опишем подход к созданию моделей на основе термов, *в качестве которых выступают распознанные слова*, этот подход будет учитывать ошибки и особенности распознавания.

Термом, извлеченным из одного или нескольких документов, используемых для построения модели, является:

$$W = (t(W), m_1(W), m_2(W), m_3(W), m_4(W), m_5(W), m_6(W)),$$

где $t(W)$ – ядро терма, то есть последовательность символов распознанного и нормализованного слова, состоящая из символов алфавита распознавания и знаков «?» и «*». Последние используются для задания множества слов, например, «*ab?c» задает множество слов с произвольным числом символов, с последними символами ab?c, на месте «?» может присутствовать произвольный символ:

$m_1(W)$ – порог при сравнении двух слов $t(W)$ и W^r. Для сравнения двух слов мы использовали расстояние Левенштейна, для слов с знаками «?» соответствующие символы в сравнении не участвуют, а для слов со знаками «*» сравниваются нужные подстроки. Если $d(t(W), W^r)<m_1(W)$, то слово W^r и терм W являются идентичными, в противном случае – различающимися. В простейшем случае $m_1(W)$ – это максимальное число операций замены при трансформации $t(W)$ в W_r;

$m_2(W)$ – ограничение на длину слова W_r при сравнении $t(W)$ и W_r, имеет смысл для слов, содержащих «*»;

$m_3(W)$ – признак зависимости от регистра (case sensitive/insensitive) при сравнении символов;

$m_4(W)$ – прямоугольник слова, состоящий из нормализованных координат $m_{4x1}(W)$, $m_{4y1}(W)$, $m_{4x2}(W)$, $m_{4y2}(W)$ в диапазоне [0,1]. Для моделей сложно структурированных документов прямоугольник слова может занимать площадь, значительно превышающую площадь конкретного слова в документе, использованного при формировании модели. В предельном случае $m_{4x1}(W)=m_{4y1}(W)=0$ и $m_{4x2}(W)=m_{4y2}(W)=1$, то есть прямоугольник слова не отличается от всего документа;

$m_5(W)$ – признак запрещенного слова, который не может присутствовать в документе. С помощью $m_5(W)$ можно организовать проверку как наличия, так и отсутствия слова W в тексте;

$m_6(W)$ – будет описан ниже.

Рис. 2. Пример зашумленного изображения и результатов его распознавания

Используется общий для всех моделей словарь $T=\{t(W_1), t(W_2), \dots t(Wz)\}$, где W_1, W_2, … Wz – набор термов, входящих хотя бы в одну модель.

При сравнении терма W со словом W^r, принадлежащим распознанному документу D, используется ядро распознанного слова $t(W^r)$ и прямоугольник слова $F(W^r)$. Сравнение терма W и слова W_r основано на условии:

$$d(t(W), W^r)<m_1(W) \wedge (F(W^r) \cap m_4(W) = F(W^r)), \quad (4)$$

где d – функция расстояния между двумя словами. При сравнении в случае case insensitive, то есть при $m_3(W)$=TRUE, оба слова $t(W)$ и W^r приводятся к одном регистру. Расстояние между термом W и словом W^r определяется как $d(t(W), W^r)$.

Для жесткоструктурированных документов при сравнении W и W_r может быть применено дополнительное условие:

$$\rho(\{F_x(W), F_y(W)\}, \{m_{4x}(W), m_{4y}(W)\})<\varepsilon, \quad (5)$$

где ρ – функция расстояния между точками $\{F_{x1}(W), F_{y1}(W)\}$ и $\{m_{4x1}(W), m_{4y1}(W)\}$, ε – предельное расстояние.

Определим для случая $m_5(W)$=0 предикат P(W, D)=1, если в тексте распознанного документа D найдено хотя бы одно слово W^r, идентичное слову W, что будем обозначать как $W^r \leftrightarrow W$, и P(W, D)=0, если в тексте не найдено ни одного слова, идентичного слову W. Если слово обладало признаком $m_5(W)$=1, то P(W, D)=0, если в тексте распознанного документа D найдено хотя бы одно слово, идентичное слову W, и P(W, D)=1, если в тексте не найдено ни одного слова, идентичного терму W.

Описанные термы W могут быть представлены как текстовые особые точки, координатами которых являются $\{m_{4x}(W), m_{4y}(W)\}$, а дескриптором – $(t(W), m_1(W), m_2(W), m_3(W), m_{4w}(W), m_{4l}(W)\}$. Слово W^r, принадлежащим распознанному документу D, также может рассматриваться как особая точка, координатами которой являются $\{F_x(W), F_y(W)\}$ а дескриптором – $(t(W^r), F_w(W), F_l(W)\}$.

Поиск соответствия между текстовой особой точкой и особой точкой в документе-запросе проводится перебором слов документа, входящих в словарь $W_r \in T$ и удовлетворяющих условиям (4) и (5). Вообще говоря, в документе D может быть найдено несколько слов $\{W^r\}$, соответствующих одной текстовой особой точке модели.

Далее определим несколько форм над несколькими термами.

Определим размещение термов как упорядоченное множество термов $R = W_1, W_2, \dots$, для которого проверятся наличие каждого из слов в распознанном документе D:

$$\mathrm{P}(W_1, T) \wedge \mathrm{P}(W_2, T) \wedge \dots \quad (6)$$

и дополнительно для каждого битерма W_i и W_{i+1} проверяется условие

$$\sigma(W^r_{i+1}, W^r_i)< m_6(W), \quad (7)$$

где параметр $m_6(W)$ ограничивает расстояние между соседними словами $W^r_i \leftrightarrow W_i$ и $W^r_{i+1} \leftrightarrow W_{i+1}$ в размещении, при $m_6(W_{i+1})=\infty$ условие (7) не проверяется, а проверяется только порядок следования слов. В качестве σ в простейшем случае применялось количество слов между W^r_i и W^r_{i+1} в линейном представлении результатов распознавания документа D, более точный способ основан на разбиении текста на параграфы.

Для размещения R может быть задан $m_7(R)$ – прямоугольник размещения, смысл которого аналогичен прямоугольнику терма, проверяется, содержатся ли полностью найденные в тексте D слова, идентичные термам $W_1, W_2 \dots$, в прямоугольнике $m_7(R)$.

Совместное выполнение условий (6), (7) и условия проверки соответствия рамке $m_7(R)$ прямоугольника определяет предикат принадлежности размещения R документу D: P(R, D)=1. В простейшем случае размещение может состоять из одного единственного терма, в общем случае для вычисления P(R, D) требуется перебор множеств, идентичных термам $W_1, W_2 \dots$. Алгоритм перебора слов распознанного D документа при определении соответствия размещению термов основан на быстром поиске в массиве слов.

Определим оценку $d(R, D)$ расстояния между размещением R и документом D как минимальную из оценок соответствия термов:

$$d(R, D)=\min(d(W_1, D),\ d(W_2, D), \dots).$$

Далее определим сочетание как множество размещений $S = R_1, R_2, \dots$, для которого проверятся наличие каждого из размещений в распознанном документе T:

$$\mathrm{P}(R_1, D) \wedge \mathrm{P}(R_2, D) \wedge \dots \quad (8)$$

Порядок размещений неважен, размещение аналогично модели «мешок слов». В дополнение к условию (8) может быть добавлено условие сравнения прямоугольников всех термов, входящих в сочетание, с прямоугольником сочетания $m_8(S)$. Указанные условия определяют предикат принадлежности сочетания S документу D: P(S, D)=1.

Определим оценку $d(S, D)$ расстояния между сочетанием S и документом D как минимальную из оценок размещений $R_1, R_2, \dots$, входящих в сочетание:

$$d(S, D)=\min(d(R_1, D),\ d(R_2, D), \dots).$$

И, наконец, определим модель M как множество сочетаний S_1, S_2,…, для которого соответствие распознанному документу D устанавливается условием:

$$\mathrm{P}(M, T) = \mathrm{P}(S_1, T) \vee \mathrm{P}(S_2, T) \vee \ldots \qquad (9)$$

В дополнение к условию (9) может быть добавлено сравнение прямоугольников всех термов модели с прямоугольником модели $m_9(M)$.

Определим оценку $r(M, D)$ расстояния между моделью и документом D как максимальную из оценок сочетаний, входящих в модель:

$$r(M, D) = \max(d(S_1, D),\ d(S_2, D), \ldots).$$

Задача выбора наилучшего класса для распознанного документа D решается ранжированием моделей M_1, …, M_n, то есть вычислением расстояний $r(M_1, D)$, …, $r(M_n, D)$, отбрасыванием таких классов M_j, что справедливо $r(M_j, D) > t_{\min}$, упорядочиванием получившегося набора по возрастанию и сохранением одной или нескольких альтернатив $r(M_{i1}, D)$, $\mathrm{r}(M_{i2}, D)$,…. Разрешение конфликтов $r(M_{i1}, D) = r(M_{i2}, D)$ в простейшем случае провести, отказавшись от классификации, в некоторых случаях конфликт устраняется использованием дополнительных признаков m_9, m_{10}.

Поясним особенности предложенных моделей.

Простые единичные ошибки распознавания вида E_4, часто появляющееся в некотором слове, могут быть проигнорированы с помощью знаков «?» и «*». Для контроля длины слова, включающего знак «*», может быть использован параметр m_2, например, ключевые слова «ДОГОВОР», «Договора» могут быть описаны ядром «ДОГОВОР*» с ограничением m_2=8, что исключает из рассмотрения слова типа «договороспособность». Ядро «?ОГОВОР» позволяет не различать слова «ДОГОВОР» и «АОГОВОР».

Порог m_1 может быть использован для требования полного совпадения ядра терма и слова из текста, например, при m_1=0 ядро «ДОГОВОР» при поиске в тексте не допускает выбора слов «АОГОВОР» или «ДОГО8ОР», отличающихся одной операцией замены символа.

Параметр m_3 также позволяет игнорировать ошибки распознавания регистра символа.

Прямоугольник слова (размещения, сочетания, модели) обеспечивает частичное описание в модели структуры документа за счет извлечения слов из указанных областей образа распознанного документа. Применение условия (5) ужесточает сопоставление особых текстовых точек и особых точек распознанного документа, что может использоваться для классификации жесткоструктурированных документов.

Ошибки вида E_3, то есть неверно распознанная структура страницы, могут быть иногда проигнорированы настройкой размещений. Размещение естественным образом описывает словосочетание или несколько расположенных рядом слов. В случае разбиения колонки текста на два столбца из-за ошибок распознавания описанное размещение будет найдено, если указывать расстояния между словами m_6=∞.

Возможность проверять как наличие, так и отсутствие слов, устанавливаемое параметром m_5, позволяет составлять модели для похожих документов, разделяя их с помощью термов, присутствующих в одном классе документов и отсутствующих в другом.

Приведенные пояснения показывают полезность описанной схемы построения моделей для классификации распознанных с ошибками текстов. Описанный способ позволяет формировать интуитивно понятные описания моделей. Например, документ «Договор аренды нежилого помещения» можно описать термами, приведенными в табл. 1. Для этого документа модель $M=S_1|S_2|S_3$ с рамкой $m_8(M)$ = {(0.0,0.0), (0.3,0.9)} состоит из сочетаний $S_1 = R_1 \wedge R_2$, $S_2 = R_1 \& R_3$, $S_3 = R_3$, для размещений $R_1=W_1$, $R_2=W_2 \otimes W_3 \otimes W_4$, $R_3=W_2 \otimes W_5 \otimes W_6$, $R_4=W_7 \otimes W_8 \otimes W_9 \otimes W_{10}$, рамки у всех термов отсутствуют.

3. Построение моделей

Опишем способ построения описанных моделей.

Рассмотрим процесс формирования моделей на реальном примере для потока документов 45 классов. Модели готовились в несколько этапов, ориентируясь на термы с первой страницы многостраничных документов.

Этап 1. Вначале рассматривалось множество эталонных документов. Для каждого класса было подготовлено несколько образцов идеальных документов, в результатах распознавания которых ошибки отсутствуют. Текстовые представления этих документов были преобразованы в мешки слов, т.е. в мультимножества слов, из числа которых были удалены стоп-слова (короткие слова, ФИО, числовые данные, даты и т.п.) и проведена текстовая нормализация. На построенных термах формируются тематические модели. Основное внимание обращалось на характерные слова из заголовков и названий разделов документа. Таким образом было составлено по нескольку размещений, хорошо отделяющих одни классы от других. То

Табл. 1

Пример термов модели документа «Договор аренды нежилого помещения»

Терм	Ядро	m_1	m_2	m_3	m_4	m_5
W_1	Договор	2	∞	1	-	∞
W_2	аренды	2	∞	0	-	∞
W_3	нежилого	2	∞	0	-	2
W_4	помещения	2	∞	0	-	2
W_5	нежилых	2	∞	0	-	2
W_6	помещений	2	∞	0	-	2
W_7	ДОГОВОР	2	∞	1	-	∞
W_8	аренд*	2	6	0	-	∞
W_9	не???лого	1	∞	0	-	2
W_{10}	Фонда	1	∞	0	-	1

есть был проведен поиск ядер термов. Далее были проанализированы проблемные классы, например, первая страница документов класса «Устав» могла быть представлена одним единственным ключевым словом «Устав», которое часто встречается в других документах, например, в документах класса «Договор». Для различения таких классов могут использоваться запрещенные слова.

Этап 2. Далее использовалась выборка документов, состоящей из 50-100 образцов реальных документов (одностраничных и многостраничных), распознавание которых давало широкий спектр ошибок всех видов $E_2 – E_4$. Анализ ошибок позволил выявить некоторое количество страниц, которые требовали модификации моделей, и страниц, которые не могли быть классифицированы из-за большого количества ошибок распознавания. Модификация моделей сводилась к поиску новых размещений и сочетаний, поиску запрещенных слов, а также к выбору параметров термов, прежде всего, рамок m_4 и расстояний между словами m_6.

Классификация выборки объемом q оценивались следующими значениями:

- n_1 – количество первых страниц документов, которые были классифицированы правильно;
- n_2 – количество первых страниц документов, которые были классифицированы неправильно;
- n_3 – количество первых страниц документов, которые не были классифицированы;
- k_1 – количество непервых страниц документов, которые не были классифицированы;
- k_2 – количество непервых страниц документов, которые были классифицированы неправильно.

Анализ результатов классификации проводился с помощью следующих критериев:

- точность классификации $a_c = (n_1+k_1)/q$;
- доля ложной классификации $z_{PF} = n_2/q$;
- доля ложной классификации $z_{NF} = n_3/q$.

Тяжелой ошибкой классификации являлась ложная классификация непервой (промежуточной и последней) страницы многостраничного документа, что объясняется сложностью поиска отсутствующей части документа в системе хранении.

Этап 3. Первые два этапа основаны на использовании правил (собственно моделей) для отнесения к тому или иному классу. Для выборок большого объема (3000 – 30000 образцов каждого класса) возможно проведение машинного обучения, например, известным методом построения бинарного дерева решений CART (Classification and Regression Trees [13]). Исходными данными для обучения являются описания известных по предыдущим этапам ключевых слов в виде ядра и признаков. На этих признаках строилось несколько деревьев для каждого класса, по стратегии один против всех. Если некоторый документ в процессе классификации не был распознан ни одним из деревьев, формируется отказ классификации. Из известных особенностей метода CART отметим необходимость достаточно большого объема обучающей выборки для стабильного обучения. Из-за этого мы не будем приводить результаты классификации, полученные с помощью метода CART, хотя в целом они превосходят результаты классификации, полученные с помощью обучения на Этапах 1 и 2.

4. Эксперименты

Для классификации потока документов, состоящего из 40 классов, использовались два множества:

Табл. 2

Результаты классификации для множества $\aleph_1$

Вариант форм	n_1	n_2	n_3	k_1	k_2	a_c	z_{PF}	z_{NF}
без рамок	773	21	229	736	9	85,35%	1,19%	0,51%
с рамками	768	13	242	743	2	85,46%	0,74%	0,11%

Табл. 3

Результаты классификации для множества $\aleph_2$

Вариант форм	n_1	n_2	n_3	k_1	k_2	a_c	z_{PF}	z_{NF}
без рамок	825	13	148	1992	36	93,46%	0,43%	1,19%
с рамками	837	1	148	2027	1	95,02%	0,03%	0,03%

– $\aleph_1$ – множество, состоящее из образов документов среднего и плохого качества оцифровки, подобранные для этапа обучения (1768 страниц);
– $\aleph_2$ – множество, состоящее из образов документов среднего качества оцифровки, полученные независимо от этапа обучения (3014 страниц).

Рассматривались два варианта работы предложенного алгоритма. В первом варианте использовались формы, включающие термы без геометрических характеристик, то есть без рамки m_4, во втором – некоторые термы обладали рамками m_4. Такое разделение должно было показать различие между применением известных структур, не обладающих геометрическими характеристиками, и структурами, использующими геометрические характеристики, полученные в процессе распознавания.

Полученные при тестировании алгоритма классификации приведены в табл. 2 и 3.

Из приведенных данных следует, что описанный метод классификации дает точность 0,86 – 0,95, при этом ложная классификация не превышает 0,01, остальные ошибки относятся к отказам от классификации. То есть предложенный метод не всегда срабатывает, но редко предлагает неверный класс.

Данные таблиц демонстрируют увеличение точности классификации и снижение долей ложной классификации и отказа от классификации за счет применения геометрических характеристик. Обратим внимание, что за счет применения рамок термов доля ложной классификации z_{NF} уменьшилась в 36 раз на множестве $\aleph_2$.

Точность полученного метода сопоставима с точностью алгоритмов, упомянутых выше [1, 4, 7].

Выводы

Описанный метод, основанный на применении текстовых особых точек, является эффективным для классификации распознанных одностраничных и многостраничных документов. Особые текстовые точки позволяют строить набор моделей, классифицирующих изображения среднего качества жесткоструктурированных и сложноструктурированных документов с высокой точностью (до 95%) и с высоким быстродействием (от 20 мсек для процессора Intel(R) Core(TM) i7).

Существенным для описанного алгоритма является применение текстовых особых точек с геометрическими характеристики, полученными при распознавании образов страниц документов. Проведенные эксперименты показывают, что классификация распознанных документов позволяет повысить точность на 1,5% по отношению к аналогичным текстам на естественно языке.

Литература

1. *Awal A.M., Ghanmi N., Sicre R., Furon T.* Complex Document Classification and Localization Application on Identity Document Images // Proc. 14th IAPR International Conference on Document Analysis and Recognition. – 2017. – P. 427-432. doi 10.1109/ICDAR.2017.77
2. *Ondrej Chum, Jiri Matas and Josef Kittler.* “Locally Optimized RANSAC”. In: DAGM-Symposium. Vol. 2781. Lecture Notes in Computer Science. 2003, P. 236–243
3. *Шемякина Ю.А., Жуковский А.Е., Фараджев И.А.* Исследование алгоритмов вычисления проективного преобразования в задаче наведения на планарный объект по особым точкам // Искусственный интеллект и принятие решений, № 1. 2017. С. 43-49.
4. *Rusiñol M., Frinken V., Karatzas D., Bagdanov A.D., Lladós J.* Multimodal page classification inadministrative document image streams // In: IJDAR 17.4 (2014), pp. 331–341.
5. *Rubin T.N.,Chambers A., Smyth P., Steyvers M.* Statistical topic models for multi-label document

classification // Machine Learning. – 2012. – Vol.88,no.1-2. – P.157208.
6. *Zhou S., Li K., Liu Y.* Text categorization based on topic model//International Journal of Computational Intelligence Systems. – 2009. – Vol.2, no.4. – P.398409
7. *Vorontsov K.V., Potapenko A.A.* Tutorial on probabilistic topic modeling: Additive regularization for stochastic matrix factorization // AIST'2014, AnalysisofImages, Social networks and Texts.- Vol.436. – Springer International Publishing Switzerland, Communications in Computer and Information Science (CCIS), 2014. – P.29-46.
8. *Воронцов К.В.* Аддитивная регуляризация тематических моделей коллекций текстовых документов // Доклады РАН. 2014. Т. 456, № 3. С. 268-271.
9. *El-Kishky A., Song Y., Wang C., Voss C. R., Han J.* Scalable topical phrase mining from text corpora // Proc. VLDB Endowment. — 2014. — Vol. 8, no. 3. — Pp. 305-316.
10. *Liu J., Shang J., Wang C., Ren X., Han J.* Mining quality phrases from massive text corpora // Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data. — SIGMOD 45. — New York, NY, USA: ACM, 2015. Pp. 1729-1744.
11. *Yarn X., Guo J., Lan Y., Cheng X.* A biterm topic model for short texts // Proceedings of the 22Nd International Conference on World Wide Web. — WWW '13.- Republic and Canton of Geneva, Switzerland: International World Wide Web Conferences Steering Committee, 2013.- P. 1445-1456.
12. *Смирнов С.В.* Технология и система автоматической корректировки результатов при распознавании архивных документов. Диссертация на соискание ученой степени кандидата технических наук, СПт:, 2015. – 130 с.
13. Breiman L., Friedman J. H., Olshen R. A., & Stone C. J. Classification and regression trees. Monterey // CA: Wadsworth & Brooks/Cole Advanced Books & Software, 1984. – 368 p.

Славин Олег Анатольевич. Институт системного анализа Федерального исследовательского центра «Информатика и управление» Российской академии наук, г. Москва, Россия. Главный научный сотрудник, доктор технических наук. Количество печатных работ: 75 (в т.ч. 1 монография). Область научных интересов: распознавание образов, информационные системы. E-mail: oslavin@isa.ru

Арлазаров Владимир Львович. Институт системного анализа Федерального исследовательского центра «Информатика и управление» Российской академии наук, г. Москва, Россия. Заведующий отделением, Чл.– корр. РАН, профессор. Количество печатных работ: более 100 статей и монографий. Область научных интересов: теория графов, распознавания образов, программирование. E-mail: arl@isa.ru

Method for classifying recognized pages of administrative documents on the basis of text key points

O.A. Slavin[I,II], V.L. Arlazarov[I,II]

[I] Institute for Systems Analysis, Federal Research Center "Computer Science and Control" of Russian Academy of Sciences, Moscow, Russia

[II] LLC "Smart Engines Service", Moscow, Russia

Abstract. The paper considers the problem of classification of recognized pages of business documents. Administrative documents used in document circulation, including in the exchange of documents between organizations, have a certain standardization, they can be both unstructured and structured. In banks or insurance companies, such documents as a power of attorney, a contract, a card with samples of signatures and seals, a charter, a contract, an account, registration certificates, etc. are often needed. When creating and maintaining electronic archives, paper documents are digitized, and digital images of pages (page scans) can be recognized and analyzed. One of the tasks of the analysis is the classification of the page image, which consists in verifying that the page image belongs to a particular class. A simple method for classifying administrative documents that yields acceptable results is proposed.

Keywords: *classification of texts; recognition of documents; OCR; recognition error; template matching.*
DOI: 10.14357/20790279180504

References

1. *Awal A.M., Ghanmi N., Sicre R., Furon T.* Complex Document Classification and Localization Application on Identity Document Images // Proc. 14th IAPR International Conference on Document Analysis and Recognition. – 2017. – P. 427-432. doi 10.1109/ICDAR.2017.77
2. *Ondrej Chum, Jiri Matas and Josef Kittler.* "Locally Optimized RANSAC". In: DAGM-Symposium. Vol. 2781. Lecture Notes in Computer Science. 2003, P. 236–243
3. *Shemyakina Y.A., Zhukovsky A.E., Faradjev I.A.* Issledovaniye algoritmov vychisleniya proyektivnogo preobrazovaniya v zadache navedeniya na planarnyy ob"yekt po osobym tochkam [Investigation of algorithms for calculating a projective transformation in the problem of targeting to a planar object from feature points], Iskusstvennyy Intellekt i Prinyatiye Resheniy [Artificial Intelligence and Decision Making], vol. 1, 2017, pp. 43-49.
4. *Rusiñol M., Frinken V., Karatzas D., Bagdanov A.D., Lladós J.* Multimodal page classification inadministrative document image streams // In: IJDAR 17.4 (2014), pp. 331–341.
5. *Rubin T.N.,Chambers A., Smyth P., Steyvers M.* Statistical topic models for multi-label document classification // Machine Learning. – 2012. – Vol.88,no.1-2. – P.157208.
6. *Zhou S., Li K., Liu Y.* Text categorization based on topic model//International Journal of Computational Intelligence Systems. – 2009. – Vol.2, no.4. – P.398409
7. *Vorontsov K.V., Potapenko A.A.* Tutorial on probabilistic topic modeling: Additive regularization for stochastic matrix factorization // AIST'2014, AnalysisofImages, Social networks and Texts.- Vol.436. – Springer International Publishing Switzerland, Communications in Computer and Information Science (CCIS), 2014. – P.29-46.
8. *Vorontsov K.V.* Additive regularization of thematic models of collections of text documents // Doklady RAS. 2014. V. 456, № 3. P. 268-271.
9. *El-Kishky A., Song Y., Wang C., Voss C. R., Han J.* Scalable topical phrase mining from text corpora // Proc. VLDB Endowment. — 2014. — Vol. 8, no. 3. — Pp. 305-316.
10. *Liu J., Shang J., Wang C., Ren X., Han J.* Mining quality phrases from massive text corpora // Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data. — SIGMOD 45. — New York, NY, USA: ACM, 2015. Pp. 1729-1744.
11. *Yarn X., Guo J., Lan Y., Cheng X.* A biterm topic model for short texts // Proceedings of the 22Nd International Conference on World Wide Web. — WWW '13.- Republic and Canton of Geneva, Switzerland: International World Wide Web Conferences Steering Committee, 2013.- P. 1445-1456.
12. Smirnov S.V. Technology and system of automatic adjustment of results under recognition of archival documents. Dissertation for the degree of candidate of technical sciences, Spb:, 2015. – 130 P.
13. *Breiman L., Friedman J.H., Olshen R.A. & Stone C.J.* Classification and regression trees. Monterey // CA: Wadsworth & Brooks/Cole Advanced Books & Software, 1984. – 368 p.

O.A. Slavin. Institute for Systems Analysis, Federal Research Center "Computer Science and Control" of Russian Academy of Sciences, Moscow, Russia. Lead researcher. Doctor of Technical Sciences. Number of publications: 75 papers, 1 book. Scientific interests: artificial intelligence, machine learning, recognition systems, information technology. E-mail: oslavin@isa.ru

V.L. Arlazarov. Institute for Systems Analysis, Federal Research Center "Computer Science and Control" of Russian Academy of Sciences, Moscow, Russia. Head of the department, Corresponding member. RAS, Professor. Number of publications: great than 100 papers and books. Scientific interests: artificial intelligence, recognition systems, information technology, programming. E-mail: arl@isa.ru