Классификация распознанных страниц деловых документов на основе метода template matching

О. А. Славин
Федеральный исследовательский центр «Информатика и управление» РАН
oslavin@isa.ru

***Аннотация*. В работе рассматривается задача классификации распознанных страниц деловых сложноструктурированных документов, состоящая в проверке принадлежности образа страницы определенному классу. К деловым относятся такие документы как доверенность, договор, карточка с образцами подписей и печатей, устав, контракт, счет, свидетельства о регистрации и т.п. Описан простой метод классификации деловых документов, дающий приемлемые результаты по точности классификации, необходимой для создании и ведении электронных архивов отсканированных бумажных документов. Предложенный метод классификации может быть применен в современных САПР для анализа содержимого текстовых документов.**

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

I. Постановка задачи

Задача классификации состоит в установлению соответствия произвольного объекта к одному или нескольким классам из заранее определенного множества классов. Алгоритм классификации может быть построен на основе машинного обучения, использующего некоторую обучающую выборку объектов, про которые известно, к каким классам они относятся.

Хорошо известен метод Template Matching (сопоставления шаблонов), состоящий в предварительной подготовке шаблонов для всех возможных классов. Принятие решения о принадлежности тестового объекта к определенному классу осуществляются по критерию минимума (максимума) некоторой функции сопоставления объекта и его шаблона (в работе [1] рассматриваются объекты – изображения символов). Классификация Template Matching основана на одной из самых простых моделей, при реализации основным является вопрос о представлении объекта некоторым шаблоном и выборе функции сопоставления.

Для анализа информации, содержащейся в образе документа (в частности, классификации) можно использовать текстовое представление документа, полученного из отсканированного образа бумажного документа программой распознавания произвольного текста (OCR).

Работа выполнена при финансовой поддержке РФФИ, проект № 16-07-01051

В настоящее время одним из самых популярных решений является OCR Tesseract, например, в работе [2] эта OCR была выбрана для распознавания корпуса архивных документов по следующим причинам:

- возможность свободного распространения;
- представление результатов распознавания в формате HOCR (HTML OCR) с сохранением информации о координатах распознанных слов.

Характерные ошибки OCR Tesseract можно разделить на несколько типов:

- E_1: полный отказ от распознавания страницы;
- E_2: число ошибок столь велико, что человек не может понять документ;
- E_3: неверно распознана структура страницы;
- E_4: наличие небольшого числа ошибок (в неверно распознанных словах присутствует 1–2 ошибки).

Рассмотрим задачу классификация образов страниц многостраничных документов. В имеющемся потоке изображений страниц одностраничных или многостраничных документов, необходимо определить границы каждого из имеющихся документов и отнести найденные документы к одному из известных заранее классов. Полученные последовательности страниц сохраняются в электронном архиве как отдельные документы.

В работе описан простой метод классификации страниц документов, основанный на сопоставлении с заранее подготовленными шаблонами классов.

Предлагается создавать шаблоны классов на основе ключевых слов (фраз) и комбинирования упорядоченных последовательностей ключевых слов (фраз).

В настоящее время известно несколько подходов к описанию моделей документа, например:

- в виде векторной модели или «мешка слов», т.е. мультимножества входящих в документ слов [3];
- в виде языковой модели, учитывающей порядок слов (учитывающей вероятность появления последовательности слов) [2].

II. Описание шаблонов и сопоставления с шаблонами

Опишем подход к созданию шаблонов, использующий перечисленные механизмы и учитывающий ошибки распознавания. В качестве признаков будут выступать распознанные слова в форме последовательностей символов и его свойств: $W=(T(W), m_1(W), \ldots, m_6(W))$, где

- $T(W)$ – *ядро слова*, то есть последовательность символов и знаков "?" и "*" (символы "?" и "*" используются для задания множества слов с произвольными символами);
- $m_1(W)$ – порог расстояния при *сравнении* ядра шаблона $T(W)$ и анализируемого W_r. Для сравнения этих двух слов предлагается использовать упрощенное расстояние Левенштейна, учитывающее только количество операций замены одного символа на другой в словах одинаковой длины. Если $d(T(W), W_r)<m_1(W)$, то слова считаются *идентичными*, в противном случае – различными;
- $m_2(W)$ – ограничение на длину слова W_r при сравнении $T(W)$ и W_r;
- $m_3(W)$ – зависимость от регистра;
- $m_4(W)$ – *рамка слова*, состоящий из координат прямоугольника, ограничивающего зону на изображении документа, в которой может размещаться слово;
- $m_5(W)$ – признак *отрицания слова*, указывающий, что данного слова не должно быть в тексте.

При поиске слова W в тексте распознанного документа T_r мы проверяем истинность следующего условия:

$$\exists W_r \in T_r : d(T(W), W_r) < m_1(W) \wedge (F(W_r) \cap m_4(W) = F(W_r)),$$

где $F(W_r)$ – рамка слова W_r, то есть координаты слова, извлеченные из результатов распознавания в формате HOCR, абсциссы и ординаты нормируются на ширину и высоту исходного изображения. Вообще говоря, может быть найдено несколько слов $\{W_r\}$, идентичных ключевому слову $T(W)$.

Определим для случая $m_5(W)=0$ предикат $P(W,T_r)=1$, если в тексте распознанного документа T найдено хотя бы одно слово, идентичное слову W, и $P(W,T_r)=0$ в противном случае. Для слов с $m_5(W)=1$, осуществляется проверка отсутствия этого слова в тексте распознанного документа.

Определим *размещение* слов как упорядоченное множество слов $R = W_1,W_2,\ldots$, для которого проверятся наличие каждого из слов в распознанном документе T_r:

$$P(W_1, T_r) \wedge P(W_2, T_r) \wedge \ldots \quad (1)$$

и дополнительно для каждой пары слов W_i W_{i+1} проверяется условие

$$r(W_{i+1}) - r(W_i) < m_6(W), \quad (2)$$

где функция $r(W)$ дает номер слова W в тексте T_r, в предположении, что все распознанные слова упорядоченном механизмами OCR. То есть параметр $m_6(W)$ определяет расстояние между соседними словами в размещении, при $m_6(W_{i+1})=\infty$ проверяется только порядок следования слов.

Выполнение условий (1), (2) определяет предикат принадлежности размещения R тексту T_r: $P(R, T_r)=1$. В общем случае для его вычисления требуется перебор множеств, идентичных словам $W_1,W_2\ldots$.

Оценку соответствия распознанному тексту T_r размещения $d(R, T_r)$ определим как

$$\min(d(W_1, T_r), d(W_2, T_r), \ldots).$$

Определим *сочетание* как множество размещений $S = R_1,R_2,\ldots$, для которого проверятся наличие каждого из размещений в распознанном документе T_r:

$$P(R_1, T_r) \wedge P(R_2, T_r) \wedge \ldots \quad (3)$$

Порядок размещений неважен.

Оценку соответствия распознанному тексту T_r сочетания $d(S, T_r)$ определим как

$$\min(d(R_1, T_r), d(R_2, T_r), \ldots).$$

И, наконец, определим *шаблон* M как множество сочетаний S_1, S_2, …, для которого принадлежность шаблона распознанному тексту T_r устанавливается проверкой выражения

$$P(M, T_r) = P(S_1, T_r) \vee P(S_2, T_r) \vee \ldots \quad (4)$$

Оценку соответствия распознанному тексту T_r шаблона $d(M, T_r)$ определим как

$$\max(d(S_1, T_r), d(S_2, T_r), \ldots).$$

В дополнение к условиям (1), (2), (3), (4) может быть добавлена проверка попадания слов размещения, сочетания или шаблона в некоторую рамку.

К имеющимся сравнениям шаблона M с распознанным текстом T_r добавим проверки соответствия некоторых свойств текста (количество символов на странице, количество колонок текста) с аналогичными свойствами шаблона.

Если для n классов задан набор шаблонов $M_1,\ldots, M_n$, то задача проверки соответствия классу M_i распознанной страницы документа T_r сводится к вычислению расстояния $d(M_i,T_r)$ и сравнению этого расстояния с заранее известным порогом d_1.

Описанные элементы шаблонов позволяют описать ключевые слова и фразы в предположении того, что в тексте T_r возможны ошибки распознавания типов E_3 и E_4. Для этого в ключевые слова и фразы задаются масками, содержащими произвольные одиночные символы и последовательности символов, а также в шаблон добавляются однотипные размещения, содержащие слова с характерными ошибками распознавания.

Задача выбора наилучшего класса решается вычислением расстояний $d(M_1,T_r),\ldots,d(M_n, T_r)$, отбрасыванием классов M_j, для которых $d(M_j,T_r)>d_1$, упорядочиванием получившегося набора по возрастанию и

сохранением одной или нескольких альтернатив $d(M_{i1},T_r),d(M_{i2},T_r),\ldots$. Разрешение конфликтов $d(M_{i1}, T_r) = d(M_{i2}, T_r)$ в простейшем случае сводится к отказу от классификации, в некоторых случаях конфликт устраняется использованием дополнительных признаков.

III. Методика обучения

Рассмотрим процесс формирования шаблонов на реальном примере для потока документов 45 классов. Шаблоны формировались в несколько этапов.

A. *Этап 1*

Вначале рассматривалось множество *эталонных документов*. Для каждого класса были подготовлены несколько образцов идеальных документов, распознанных без ошибок. Текстовые представления этих документов были преобразованы в мешки слов и мешки фраз, т.е. в мультимножества слов и фраз, из числа которых были удалены *стоп-слова*. Простыми переборными алгоритмами из этих мультимножеств были выбраны одиночные слова и фразы из нескольких слов, присущие одному из классов. Основное внимание обращалось на характерные слова из заголовков и названия разделов документа. Таким образом было выбрано несколько размещений, хорошо отделяющих одни классы от других.

B. *Этап 2*

Далее в выборке документов небольшого объема, состоящей из образцов *реальных документов* (одностраничных и многостраничных), распознавание которых давало широкий спектр ошибок всех видов E_1 – E_4, выбирались страницы, которые требовали модификации шаблонов, и страницы, которые не могли быть классифицированы предложенным методом (прежде всего, из-за большого количества ошибок распознавания). 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* $(n_1+k_1)/q$;
- *доля ложной классификации* $(n_2+k_2)/q$;
- *доля отказов от классификации* n_3/q, где $q=(n_1+n_2+n_3+m_1+m_2)$.

C. *Этап 3*

Первые два этапа основаны на *использовании правил* (собственно шаблонов) для отнесения к тому или иному классу. Для выборок достаточно большого объема (3000 – 30000 образцов каждого класса) возможно проведение *машинного обучения*, например, известным методом построения бинарного дерева решений CART (Classification and Regression Trees [4]).

IV. Эксперименты

Для потока документов, состоящего из 45 классов, были получены следующие результаты классификации на двух тестовых множествах:

- состоящего из образов документов среднего и плохого качества оцифровки, подобранные для этапа обучения (884 страницы) - n_1=397, n_2=3, m_2=3, n_3=111, m_1=370;
- состоящего из образов документов среднего качества оцифровки, полученные независимо от этапа обучения (3014 страниц) - n_1=832, n_2=8, m_2=1, n_3=146, m_1=2027.

Для другого потока одностраничных документов, состоящего из 1000 страниц 4 классов, были получены следующие характеристики: n_1=954, n_2=6, n_3=40 (m_1 и m_2 для одностраничных документов не подсчитывались).

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.

Точность полученного метода достаточно высока по сравнению с точностью алгоритмами классификации web-страниц, описанными в [5] (точность равна 0,75).

Быстродействие реализованного метода классификации достаточно велико: 3000 распознанных страниц обрабатываются примерно за 1 минуту на компьютере Intel(R) Core(TM) i7-4790 CPU 3.60 GHz, 16,0 GB, Windows 7 prof 64-bit.

V. Заключение

Описанный метод классификации является простым для самостоятельной реализации. Для уменьшения доли страниц, оставшихся неклассифицированными, возможны следующие шаги:

- применение методов бинаризации для снятия фона на загрязненных документа и документах со сложным фоном;
- использование более точных систем распознавания документов (OCR).

Предложенный метод позволяет классифицировать как одностраничные, так и многостраничные документы.

Предложенный метод классификации может быть применен в современных САПР, позволяющих наряду с основной функцией извлечения текстовой информации решать задачи анализа содержимого текстовых документов.

Список литературы

[1] Гонсалес Р., Вудс Р. Цифровая обработка изображений. М: Техносфера, 2005. 1070 с.

[2] Смирнов С.В. Технология и система автоматической корректировки результатов при распознавании архивных документов: дис. на соиск. учен. степ. канд. техн. наук / СПИИРАН. СПб., 2015. 130 с.

[3] Martin D., Jurafsky D. Speech and Language Processing. An introduction to natural language processing, computational linguistics, and speech recognition. New Jersey: Pearson Prentice Hall, 2009. 988 p.

[4] 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.

[5] Маслов М.Ю., Пяллинг А.А., Трифонов С.И. Автоматическая классификация веб-сайтов. RCDL, Дубна, Россия, 2008, с. 230-235. [Электронный ресурс]. – Режим доступа http://rcdl2008.jinr.ru/pdf/230_235_paper27.pdf. Дата обращения: 15.04.2018.