POLITECHNIKA WARSZAWSKA
WYDZIAŁ ELEKTRYCZNY
INSTYTUT STEROWANIA I ELEKTRONIKI PRZEMYSŁOWEJ

PRACA DYPLOMOWA MAGISTERSKA
na kierunku INFORMATYKA

Łukasz Murawski
Nr ind. 181947

Rok akad.: 2005/2006
Warszawa, 23.11.2005

Wykorzystanie wybranych metod przetwarzania obrazów do wspomagania procesu ekstrakcji treści z prasy elektronicznej

Zakres pracy:

1. Wprowadzenie
2. Wybrane aspekty budowy plików PDF
3. Opis wybranych algorytmów analizy i dekompozycji dokumentów
4. Realizacja systemu ekstrakcji danych
5. Podsumowanie i wnioski

KIEROWNIK ZAKŁADU STEROWANIA
Dr hab. inż. Bartłomiej Beliczyński

Kierujący pracą: dr inż. Witold Czajewski

Termin wykonania: 15.09.2006

Praca wykonana i zaliczona pozostaje własnością Instytutu i nie będzie zwrócona wykonawcy

Wykorzystanie wybranych metod przetwarzania obrazów do wspomagania procesu ekstrakcji treści z prasy elektronicznej

Streszczenie

Niniejsza praca poświęcona jest przetwarzaniu plików PDF, ze szczególnym uwzględnieniem elektronicznych wersji prasy codziennej zapisanej w tym właśnie formacie. W pierwszej kolejności przedstawiono budowę formatu, jego najważniejsze cechy oraz kluczowe struktury wewnętrzne. Następnie dokonano formalnego opisu zadania głównego, którym jest proces analizy dokumentu. W tym miejscu opisano aktualny dorobek naukowy w tej dziedzinie oraz przedstawiono istniejące metody i trendy.

W trzeciej, głównej części pracy opisano wyniki prowadzonych badań. W toku prac zdefiniowane zostało pojęcie *separatora wirtualnego* - jest to nieistniejąca fizycznie linia, która oddziela szpalty tekstu w artykule prasowym. Zbadano kilka metod detekcji, a dla rokującej najlepsze rezultaty, dokonano implementacji ekstraktora wspomnianych separatorów oraz wykazano, że pozyskane w ten sposób informacje znacząco przyczyniają się do zwiększenia skuteczności procesu ekstrakcji elementów artykułu prasowego. W ostatniej części pracy zaproponowano kierunki rozwoju i dalszych badań.

Zaproponowane rozwiązania cząstkowe bazują na przetwarzaniu obrazów, a cały system ma charakter hybrydowy, z uwagi na wykorzystanie natywnych cech formatu PDF.

Use of selected image processing methods in support of electronic press contents extraction

Abstract

This thesis is devoted to the processing of the PDF files, particularly electronic editions of newspapers stored in this format. In the beginning, the structure of the format, its major features and key internal structures are presented. This is followed by a formal description of the main task - the process of document analysis. A description of the state-of-the-art and existing methods and trends is also given.

The third, main part of the thesis describes the results of the conducted research. In the course of work the concept of virtual separator was defined - a physically non-existent line that separates columns of text in a newspaper article. Several identification methods were evaluated. For the most promising one, virtual separator detector was applied and demonstrated significant increase of the extraction efficiency. In the last part of the thesis possible directions of further research and development are given.

The proposed partial solutions are based on image processing, yet the entire system is a hybrid one as it relies on native features of the PDF format.

Politechnika Warszawska
Wydział Elektryczny

Warszawa, dnia 15.09.2006

OŚWIADCZENIE

Świadom odpowiedzialności prawnej oświadczam, że niniejsza praca dyplomowa magisterska pt. Wykorzystanie wybranych metod przetwarzania obrazów do wspomagania procesu ekstrakcji treści z prasy elektronicznej:

- została napisana przeze mnie samodzielnie,
- nie narusza niczyich praw autorskich,
- nie zawiera treści uzyskanych w sposób niezgodny z obowiązującymi przepisami.

Oświadczam, że przedłożona do obrony praca dyplomowa nie była wcześniej podstawą postępowania związanego z uzyskaniem dyplomu lub tytułu zawodowego w uczelni wyższej. Jestem świadom, że praca zawiera również rezultaty stanowiące własności intelektualne Politechniki Warszawskiej, które nie mogą być udostępniane innym osobom i instytucjom bez zgody Władz Wydziału Elektrycznego.

Oświadczam ponadto, że niniejsza wersja pracy jest identyczna z załączoną wersją elektroniczną.

Łukasz Murawski...Łukasz Murawski

Spis treści

Wstęp

Początek XXI w. określa się mianem *ery internetu*. Dostęp do ogólnoświatowej sieci jest coraz bardziej powszechny, najmłodszym trudno sobie wyobrazić, że jeszcze kilkanaście lat temu posiadanie *komputera* było niespełnionym marzeniem milionów Polek i Polaków, a o internecie mało kto wtedy słyszał. Obecnie sieć oplata nas z każdej strony i jest wszechobecna - w telefonie, komputerze, telewizorze czy nawet lodówce. Pojawiają się coraz to nowe pomysły na jej wykorzystanie w codziennym życiu, powstają nowe usługi czy portale społecznościowe. Internet wkracza w niemal każdą sferę życia wyznaczając nowe standardy i mało kto może oprzeć się temu działaniu. Podobnie było i z rynkiem prasowym - coraz większa rzesza czytelników rezygnuje z tradycyjnych form druku na rzecz e-papieru. Taki stan rzeczy wymusza na wydawcach dostosowanie się do potrzeb rynku, na którym coraz częściej pojawiają się elektroniczne wersje tytułów dostępnych w kioskach i księgarniach. Coraz większe zainteresowanie budzi możliwość tworzenia tematycznych baz danych, również na potrzeby public relations, wyłącznie przy użyciu komputera. Codzienny rytuał kupowania gazet i zakreślania interesujących artykułów przez sekretarkę odchodzi w niebyt. Zamiast tego coraz więcej firm i agencji PR korzysta z usług firm profesjonalnie zajmujących się monitoringiem mediów.

Efektywna analiza treści e-prasy może być prowadzona jedynie przy wykorzystaniu komputera oraz odpowiednio przygotowanego oprogramowania. Niniejsza praca wychodzi naprzeciw tym potrzebom. Podjęto bowiem próbę zwiększenia ogólnej efektywności działania znanych metod klasyfikacji i segmentacji treści, przy użyciu metod przetwarzania obrazu. Praca ma charakter badawczy, tzn. jej celem głównym jest określenie przydatności metody oraz

wskazanie możliwych kierunków prowadzenia dalszych badań, zamiast stworzenia kompletnego i gotowego do natychmiastowego wdrożenia systemu o charakterze komercyjnym.

Przyjęto, że dane elektroniczne przekazywane są w postaci pliku PDF, co aktualnie jest niekwestionowanym standardem. Format ten jeszcze do niedawna był okryty tajemnicą, z uwagi na ograniczenia licencyjne ustanowione przez właściciela - firmę Adobe. Dopiero od stycznia 2007 r. jego specyfikacja stała się ogólnie dostępna co umożliwiło bardziej wszechstronny rozwój oprogramowania wspomagającego - szczególnie typu open source. Filozofia zapisu danych w tym formacie jest na tyle nietypowa, że cały pierwszy rozdział poświęcony został omówieniu samego formatu. Zwrócono uwagę na aspekty podstawowe oraz istotne dla niniejszej pracy. W rozdziale drugim zaprezentowano podstawowe techniki analizy treści, dokonując podziału na analizę fizyczną oraz logiczną. Rozdział trzeci opisuje teoretyczne podstawy przetwarzania obrazów, w kolejnym natomiast przeprowadzono praktyczne badania polegające na stworzeniu hybrydowego systemu segmentacji treści artykułów korzystającego z opisanych metod. W ostatnim, piątym rozdziale podsumowano przeprowadzone badania oraz wskazano drogi dalszego rozwoju podjętego zagadnienia.

Rozdział 1

Format PDF

W ostatnich latach format PDF stał się faktycznym standardem. Każdego dnia korzystają z niego tak finansiści jak i archiwiści - pierwsi publikują wyniki swoich najnowszych analiz, podczas gdy drudzy umieszczają w nich wyniki skanu starodruków. Jednym słowem korzystają z niego wszyscy, którym zależy na wiernym zachowaniu i odzwierciedleniu wszystkich szczegółów *graficznych*. Korzystają z niego również różnorodne wydawnictwa, które w tej właśnie formie udostępniają w wersji elektronicznej swoje periodyki. W niniejszym rozdziale przedstawiono najważniejsze cechy formatu oraz krótki rys historyczny zmian, jakie w nim nastąpiły na przestrzeni kilku ostatnich dekad.

1.1 Historia formatu

Format PDF wywodzi się z rodziny języków opisów strony PDL[31]. Języki tej klasy pozwalają opisać zawartość w sposób niezależny od rodzaju, a nawet typu urządzenia wyjściowego, którym może być drukarka, ploter, monitor ekranowy, rzutnik, smartphone lub dowolne inne, o ile wyposażone jest w odpowiedni interpreter. Ich rozwój częściowo wynikał z postępu w dziedzinie druku laserowego, który umożliwiał uzyskanie wyników o coraz większej rozdzielczości. Szczególnie ważne stało się więc zapewnienie zgodności kompozycji, proporcji, kształtu czcionek, kolorystyki poszczególnych elementów

oraz zminimalizowanie efektu aliasingu dla osadzanych obrazów rastrowych. Najpowszechniej znanym reprezentantem grupy jest PostScript [32]. Opublikowany w roku 1982 umożliwiał realizację postawionych wyżej zadań. PostScript jest kompletnym językiem programowania opartym na architekturze stosu oraz notacji postfixowej, z możliwością tworzenia procedur i funkcji oraz deklarowania zmiennych[24]. Interesującą cechą języka jest fakt, że zarówno tekst jak i grafika są traktowane w sposób ujednolicony. Dzięki temu elementy tekstowe mogą być skalowane, wypełniane gradientem, wzorem itp. W celu uzyskania obrazu rastrowego strony, konieczne jest jedynie uruchomienie programu na urządzeniu wyjściowym. Przykładowy program znajdziemy na rysunku 1.1.

```
%!PS
/Courier              % name the desired font
20 selectfont         % choose the size in points and establish
                      % the font as the current one
72 500 moveto         % position the current point at
                      % coordinates 72, 500 (the origin is at the
                      % lower-left corner of the page)
(Hello world!) show   % stroke the text in parentheses
showpage              % print all on the page
```

Rysunek 1.1: Prosty program w języku Postscript. Źródło:[32].

Format PDF został zaprezentowany światu przez firmę Adobe na początku lat 90. ubiegłego stulecia. Według twórców miał stać się uniwersalnym i niezawodnym medium do przeglądania, drukowania oraz przekazywania informacji pochodzących z dowolnego źródła, między różnymi osobami i instytucjami. PDF wyewoluował z wcześniej opisywanego języka Postscript. O ile jednak Postscript jest językiem programowania używanym głównie przez plotery i drukarki laserowe, o tyle PDF został przystosowany do prezentacji na ekranie monitora. Istotną różnicą w tym zakresie jest umożliwienie renderowania poszczególnych stron w PDF. W przypadku Postscript wyświetlenie określonej strony wymagało uruchomienia programu i wykonania wszystkich instrukcji dla stron poprzedzających - co generuje narzut czasowy. Wynikało to również z faktu, że Postscript wykorzystuje tylko jeden, globalny kontekst wykonania dla całego dokumentu.

Mimo tego, że już pierwsza wersja formatu PDF była w pełni funkcjonalna (patrz tabela 1.1), z biegiem lat była ona uzupełniana i rozszerzana. W wer-

Wersja	Rok publikacji	Nowe funkcjonalności
1.0	1993	
1.1	1996	szyfrowanie, łączenie bloków tekstu w artykuły, przestrzenie kolorów
1.2	1996	formularze, elementy interaktywne, zdarzenia myszy, obsługa zewnętrznych klipów filmowych, obsługa zewnętrznych oraz wbudowanych klipów dźwiękowych, kompresja danych, wsparcie dla Unicode
1.3	2000	*DeviceN, ICC* nowe przestrzenie kolorów, sygnatura cyfrowa, obsługa JavaScript, nowe struktury danych
1.4	2001	wsparcie dla *JBIG2*, przeźroczystość, tagowanie treści
1.5	2003	wsparcie dla *JPEG2000*
1.6	2006	szyfrowanie *AES*, osadzanie załączników i wzajemna nawigacja między nimi
1.7	2008	

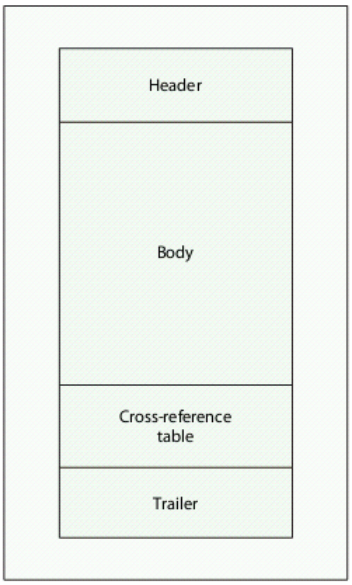
Tablica 1.1: Historia wydań PDF oraz najważniejszych funkcji. Na podstawie:[15, 31]

sji 1.2 dodano możliwość oznaczania w sposób specjalny fragmentów dokumentu oraz ich komentowania, a także tworzenia i wypełniania formularzy interaktywnych, również przez osoby niepełnosprawne. Od wersji 1.3 możliwe było dodanie interaktywności z wykorzystaniem języka JavaScript, a także uwierzytelnienie dokumentu poprzez dodanie podpisu cyfrowego. W kolejnej odsłonie umożliwiono przechowywanie w jednym pliku - dla tekstowych dokumentów skanowanych - wersji graficznej oraz tekstowej pozyskanej w wyniku działania programu rozpoznawania pisma OCR. Warstwy umieszczone nakładają się na siebie, a czytający ma możliwość regulacji transparentności, dzięki czemu może łatwo powiązać widoczne treści ze sobą. Upowszechnienie się internetu miało duży wpływ na kolejną wersję 1.5, w której dodano możliwość osadzenia hiperłączy oraz treści multimedialnych. Dane strumieniowe mogą być również pobierane na żądanie z sieci i prezentowane bezpośrednio w czytniku PDF. Taki zabieg znacząco zmniejszył objętość plików. Dalszemu rozwojowi podlegało szyfrowanie i zabezpieczenie dokumentu przed modyfikacją oraz nieuprawnionym dostępem - obecnie możliwe jest szyfrowanie przy pomocy algorytmu AES przy użyciu klucza o długości 256 bitów.

1.2 Struktura pliku i składnia

Struktura wewnętrzna pliku została podzielona na cztery główne części, jak to pokazano na rysunku 1.2.

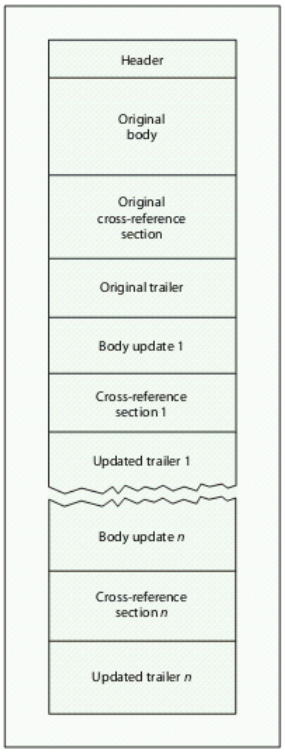
Nagłówek `header` obejmuje jedną lub dwie pierwsze linie pliku i obowiązkowo zawiera ciąg *%PDF-*, po którym następuje oznaczenie numeru wersji (*1.0,1.1 . . . 1.7*). Jeżeli plik zawiera dane binarne, co zazwyczaj ma miejsce, kolejna linia zawiera przynajmniej 4 znaki binarne tj. o kodach większych niż 127. Zabieg ten ma na celu zapewnienie poprawnej klasyfikacji pliku jako binarnego przez aplikacje służące do transferu plików (np. FTP), które zazwyczaj analizują dane w pobliżu początku pliku. W kolejnej sekcji `body` umieszczone są definicje wszystkich elementów składających się na właściwą treść dokumentu. Tabela `cross-reference table` umożliwia wydajne nawigowanie oraz bezpośredni dostęp do poszczególnych elementów. Stopka pliku `trailer` zawiera m.in. informacje o wielkości tabeli `cross-reference`,



Rysunek 1.2: Struktura pliku PDF. Na podstawie:[15].

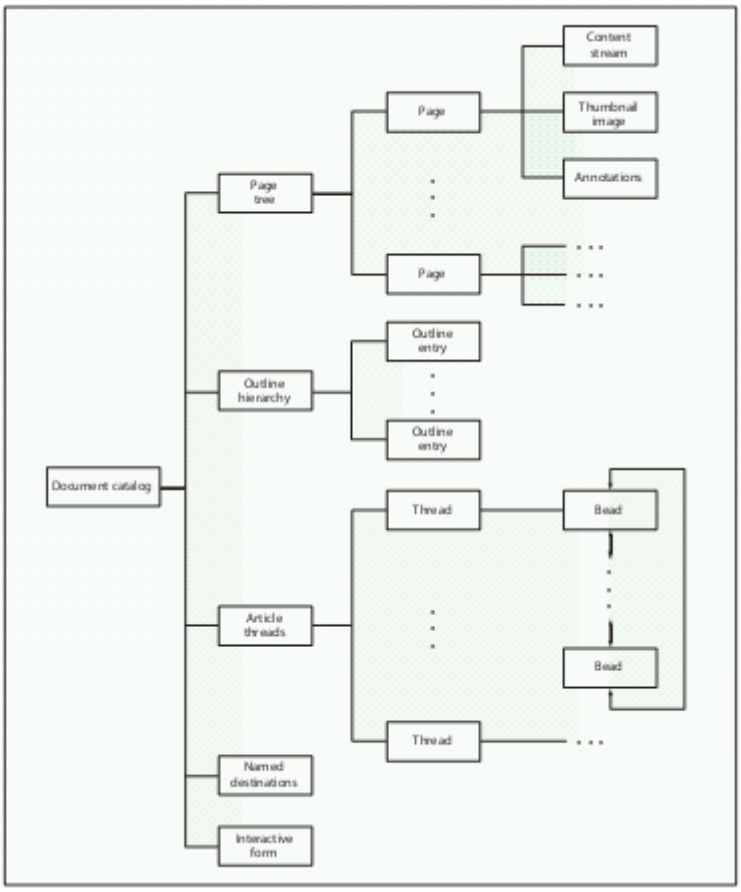
adresie poprzedniego bloku stopki (o ile jest obecny), informacje o użytym szyfrowaniu, a także offset bajtowy dla tabeli `cross-reference`. W efekcie odczyt pliku PDF praktycznie rozpoczyna się od jego końca, w celu pozyskania informacji do dalszej nawigacji. Wszystkie sekcje za wyjątkiem nagłówka mogą być aktualizowane w sposób przyrostowy, wobec czego finalnie struktura pliku może wyglądać jak na rysunku 1.3.

Zawartość pliku PDF jest osadzona w sekcji `body` pliku w sposób ustandaryzowany. Wszystkie obiekty zorganizowane są w hierarchię, a dostęp do nich możliwy poprzez unikalny numer identyfikacyjny. Obecnie obsługiwane są następujące typy danych `array, Boolean, dictionary, integer, name, null, real, stream, string`. Typy złożone jak tablica, słownik czy strumień mogą być osadzone w sposób bezpośredni lub pośredni poprzez referencję. Każdy plik PDF zawiera punkt wejścia nazywany katalogiem, który zawiera referencje do poddrzewa, którego obiektami są strony. Węzły stron są grupowane według współdzielonych lub podobnych właściwości, co ułatwia wewnętrzną nawigację w pliku, podczas gdy pojedynczy węzeł zawiera treść



Rysunek 1.3: Struktura pliku PDF w trybie aktualizacji. Na podstawie:[15].

określonej strony. Na każdą stronę wyświetlaną w czytniku PDF przypada dokładnie jeden węzeł strony. Węzeł strony składa się z kolekcji właściwości,



Rysunek 1.4: Struktura dokumentu. Na podstawie:[15].

przynależnych zasobów oraz jednego lub więcej strumieni z zawartością. Na właściwości składają się takie informacje jak rozmiar strony, jej orientacja, aktywny obszar kreślenia itp. Strumień danych stanowi ciąg bajtów tworzących określone jednostki leksykalne, które mogą być przeglądane za pomocą dowolnego edytora tekstowego.

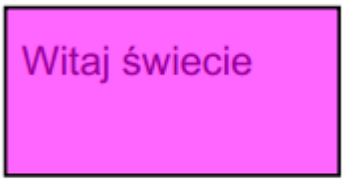
Listing 1.1: Forma źródłowa prostego dokumentu PDF.

```
%PDF-1.4
%       Ó
3 0 obj
<</Length 102/Filter/FlateDecode>>stream
%%zawartość binarna została pominięta!
endstream
endobj
5 0 obj
<</Parent 4 0 R/Contents 3 0 R/Type/Page/Resources<</ExtGState
    <</GS1 1 0 R>>/ProcSet [/PDF /Text /ImageB /ImageC /ImageI]/
    Font<</F1 2 0 R>>>>/MediaBox[0 0 595 842]>>
endobj
2 0 obj
<</LastChar 156/BaseFont/Helvetica/Type/Font/Encoding<</Type/
    Encoding/Differences[32/space 87/W 97/a 99/c 101/e 105/i/j
    116/t 119/w 156/sacute]>>/Subtype/Type1/Widths[278 0 0 0 0 0
    0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
     0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 944 0 0 0 0 0 0 0 0 0
    556 0 500 0 556 0 0 0 222 222 0 0 0 0 0 0 0 0 278 0 0 722 0
     0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
    0 0 0 0 0 500]/FirstChar 32>>
endobj
1 0 obj
<</ca 0.6>>
endobj
4 0 obj
<</ITXT(5.3.1)/Type/Pages/Count 1/Kids[5 0 R]>>
endobj
6 0 obj
<</Type/Catalog/Pages 4 0 R>>
endobj
7 0 obj
<</Producer(iText 5.3.1 2000 -2012 1T3XT BVBA \(AGPL-version
    \))/ModDate(D:20120903153417+02)/CreationDate(D
    :20120903153417+02)>>
endobj
xref
0 8
```

```
28 0000000000 65535 f
29 0000000833 00000 n
30 0000000365 00000 n
31 0000000015 00000 n
32 0000000860 00000 n
33 0000000184 00000 n
34 0000000923 00000 n
35 0000000968 00000 n
36 trailer
37 <</Root 6 0 R/ID [<4be5fe430a2a91099cbe113ccb2ad288><5
     dd0b9694c669880418af1a6e6fea4fa>]/Info 7 0 R/Size 8>>
38 startxref
39 1121
40 %%EOF
```

Na listingu 1.1 przedstawiono postać źródłową dokumentu zaprezentowanego na rysunku 1.5, na której można wyodrębnić wymienione wcześniej cztery główne składowe - nagłówek, ciało dokumentu, tabelę referencji oraz stopkę. Poszczególne obiekty odseparowane są od siebie za pomocą słów kluczowych `obj` oraz `endobj`. Definicja obiektu jest poprzedzona przez jego identyfikator, na który składają się dwie części - *object number* porządkowy numer obiektu, *generation number* mówiący o numerze wersji obiektu. Zdefiniowany obiekt może być użyty dowolną ilość razy poprzez referencję, która kreowana jest za pomocą identyfikatora i następującego po nim słowa kluczowego np. `12 0 R`. Nazwy i komentarze są odpowiednio poprzedzone znakami \ oraz `%`. Obiekty słownikowe są ograniczone za pomocą znaków $\ll$ $\gg$, zbudowane są z par klucz-wartość. Tablice - będące w istocie jednowymiarowymi kolekcjami obiektów - składać się mogą z dowolnej liczby obiektów dowolnego typu (np. dowolna kombinacja numerów, napisów, słowników i innych włączając w to inne tablice). Tablica domyślnie jest jednowymiarowa, lecz możliwe jest jej zagłębienie na dowolny poziom poprzez użycie kolejnych tablic jako elementów. Strumień danych następuje po słowie kluczowym `stream` i kończy się sekwencją `endstream`. Sam strumień poprzedzony jest słownikiem zawierającym obowiązkowo informację o długości strumienia (wartość może być podana wprost lecz zazwyczaj stosuje się odwołanie przez referencję) oraz opcjonalnie inne informacje pomocne do prawidłowej interpretacji danych

(np. o użytym algorytmie kompresji czy szyfrowania).



Rysunek 1.5: Graficzny obraz prostego dokumentu PDF

1.3 Strumień danych

Wszystkie elementy graficzne w pliku PDF (teksty, obrazy) są opisane w dedykowanym strumieniu danych `content stream`, który składa się z sekwencji instrukcji niezbędnych do odwzorowania wyglądu na urządzeniu wyjściowym (np. ekranie monitora czy podczas wydruku). Przykładową sekwencję pokazano na listingu 1.2, który zawiera dane pominięte na listingu 1.1. Po dekompresji strumienia można zauważyć, że składa się on z operandów oraz operatorów wyrażonych za pomocą znaków ASCII. W PDF, podobnie jak w PostScript, wyrażenia zapisywane są w notacji postfiksowej. Operator wraz z poprzedzającymi go operandami tworzy pojedynczą instrukcję, która opisuje renderowany element lub właściwości kontekstu graficznego.

Kontekst graficzny jest mechanizmem pozwalającym kontrolować stan *wirtualnego pióra*. Jest on modyfikowany za pomocą dedykowanych instrukcji oraz może być zapamiętywany i odtwarzany przy pomocy stosu. Wszystkie prymitywy graficzne - włączając w to kształty liter - są kreślone w aktualnym kontekście graficznym, który określa takie parametry jak macierz transformacji `CTM`, przestrzeń kolorów, kolor, kształt linii, poziom przeźroczystości i podobne. Macierz transformacji `CTM` definiuje mapowanie z układu współrzędnych użytkownika na układ współrzędnych urządzenia wyjściowego. Podczas renderowania poszczególnych elementów jest ona automatycznie modyfikowana dzięki mechanizmowi pozycjonowania relatywnego.

Listing 1.2: Zdekodowana zawartość strumienia danych (pominiętego na listingu 1.1)

```
5 0 obj
<<
/Length 7 0 R
>>
stream
q
BT
36 806 Td
0 0 Td
/F1 12 Tf
(Witaj świecie)Tj
0 0 Td
ET
Q
q
/GS1 gs
1 0 1 rg
1 w
31 776 100 50 re
B
Q

endstream
endobj
```

1.4 Reprezentacja tekstu

W większości wypadków zawartość tekstowa jest definiowana jako wektor liter. W momencie renderowania strony kreślone są glify zdefiniowane dla użytej czcionki i stają się one elementami *graficznymi*, niemniej posiadają szereg dodatkowych cech takich jak: definicja czcionki, jej wielkość, odległość między znakami czy słowami. Parametry te są zgromadzone w dedykowanej macierzy - podobnej do macierzy transformacji `CTM`. Listing 1.3 prezentuje sekwencję zaangażowaną w wygenerowanie tekstu `Witaj świecie!` pokaza-

nego na rysunku 1.5. Widoczne jest, że sekwencja ograniczona jest za pomocą operatorów `BT` oraz `ET`. Pierwsza instrukcja `Td` poprzedzona jest dwiema liczbami całkowitymi i odpowiada za modyfikację bieżących współrzędnych w macierzy `CTM`. Następnie operator `Tf` dokonuje selekcji czcionki - symbolicznie nazwanej tutaj jako `/F1` - oraz ustala rozmiar czcionki na 12 punktów. Operator Tj renderuje poprzedzający go ciąg znaków. W bardziej ogólnym wypadku w bezpośrednim sąsiedztwie znaków mogą znaleźć się instrukcje zmieniające domyślny odstęp między nimi lub same znaki mogą być zapisywane w notacji heksadecymalnej.

Listing 1.3: Fragment odpowiedzialny za renderowanie tekstu.

```
1 BT
2 36 806 Td
3 /F1 12 Tf
4 (Witaj świecie)Tj
5 ET
```

Należy podkreślić, że zawartość tekstowa w dokumencie PDF zupełnie pozbawiona jest znaczników wskazujących na logiczny podział tekstu na akapity czy rozdziały. Znaki nie muszą być grupowane w żadne większe jednostki leksykalne jak słowa czy linie. Z tego właśnie powodu analiza danych tekstowych zapisanych w formacie PDF przysparza tylu trudności.

1.5 Reprezentacja elementów graficznych

Elementy graficzne (np. obrazy wektorowe) reprezentowane są wewnętrznie w postaci ścieżek (linie lub krzywe Beziera). Każda ścieżka renderowana jest z wykorzystaniem kontekstu graficznego, przez co możliwe jest dziedziczenie i współdzielenie wartości poszczególnych atrybutów. Ścieżki używane są do rysowania obwiedni, wypełniania obszarów, opisywania czcionek itd.

Listing 1.4: Fragment odpowiedzialny za renderowanie prostokąta.

```
1 Q
2 q
3 /GS1 gs
4 1 0 1 rg
5 1 w
6 31 776 100 50 re
7 B
8 Q
```

Listing 1.4 prezentuje sekwencję instrukcji generującą prostokątne tło widoczne na obrazie 1.5. Operand `gs` odwołuje się do zdefiniowanego kontekstu graficznego, `rg` określa aktualny kolor kreślenia w przestrzeni RGB, `w` definiuje szerokość linii, `re` określa współrzędne prostokąta, a wreszcie `B` dokonuje wypełnienia obszaru podanym kolorem.

1.6 Reprezentacja obrazów rastrowych

Obrazy rastrowe w większości wypadków są osadzone w pliku PDF w postaci źródłowej - tj. binarnym strumieniu danych. Dzięki temu mogą być łatwo odnajdywane, przekształcane w aktualnym kontekście graficznym i macierzy `CTM` oraz prezentowane. Obrazy mogą być również dołączane jako obiekty zewnętrzne (w stosunku do samego pliku PDF), niemniej nie jest zalecane tworzenie dokumentu ze zbyt dużą ilością odwołań do źródeł obcych (które z czasem mogą się zdezaktualizować).

Dane binarne są zazwyczaj kompresowane przy użyciu takich algorytmów jak *CCITT* dla obrazów binarnych, *LZW* dla obrazów renderowanych czy `JPG` dla zdjęć. Różne przestrzenie barw i profile mogą zostać zdefiniowane i osadzone w pliku PDF w celu precyzyjnego odzwierciedlenia schematu barw.

1.7 Reprezentacja czcionek

PDF obsługuje różnorodne formaty czcionek włączając tzw. formaty proste, jak również czcionki kompozytowe. Czcionki proste mogą zawierać co

najwyżej 256 glifów, podczas gdy kompozytowe pozbawione są ograniczeń w tym zakresie. Niemniej - w celu zmniejszenia rozmiaru wynikowego pliku PDF - często zamiast całej czcionki w pliku osadza się jedynie te glify, które zostały użyte. Czcionki osadzone są w strumieniu danych oraz obsługiwane przez komponent odpowiedni dla jej typu. W celu poprawnego odwzorowania tekstu na ekranie przy pomocy glifów, konieczna jest obecność mechanizmu mapującego poszczególne kształty na odpowiadające im proste znaki. Możliwe jest wprowadzenie różnorodnych pośrednich mapowań dzięki mechanizmowi słowników.
W celu umożliwienia poprawnej ekstrakcji tekstów, każda czcionka powinna posiadać mapowanie do formatu UTF-8.

Rozdział 2

Techniki analizy treści

Upowszechnienie się formatu PDF jako medium przekazywania informacji oraz upublicznienie i ustandaryzowanie formatu źródłowego spowodowało pierwotnie wzrost zapotrzebowania na aplikacje i systemy do kreowania, a następnie do przetwarzania i dekompozycji dokumentów. Poznanie i rozumienie procesu tworzenia dokumentów elektronicznych jest kluczowe dla późniejszego rozumienia zagadnień dekompozycji i ekstrakcji.

2.1 Formatowanie tekstu źródłowego

W *pliku znakowym* każdy bajt (czasem grupa sąsiadujących ze sobą bajtów) zapisu informacji reprezentuje pojedynczy znak, którego postać graficzną określa zastosowany kod informacyjny (np. UNICODE czy ASCII). Wyróżnienie fragmentu tekstu (np. pogrubienie znaków tekstu) napotyka na trudności, albowiem informacje o oczekiwanym sposobie prezentacji graficznej muszą zostać zapisane w tym samym pliku, ale taka informacja - jako że zapisana w ciągu bajtów - byłaby zrozumiana nie jako opis właściwości pewnych znaków, ale jako dodatkowe znaki tekstu, o postaci określonej przez kod informacyjny [17].
Tekst sformatowany, oprócz warstwy treściowej, może zawierać dodatkowe informacje na temat swojej struktury i sposobu prezentacji. Opis struktury polega na określeniu roli poszczególnych fragmentów w dokumencie (np. ty-

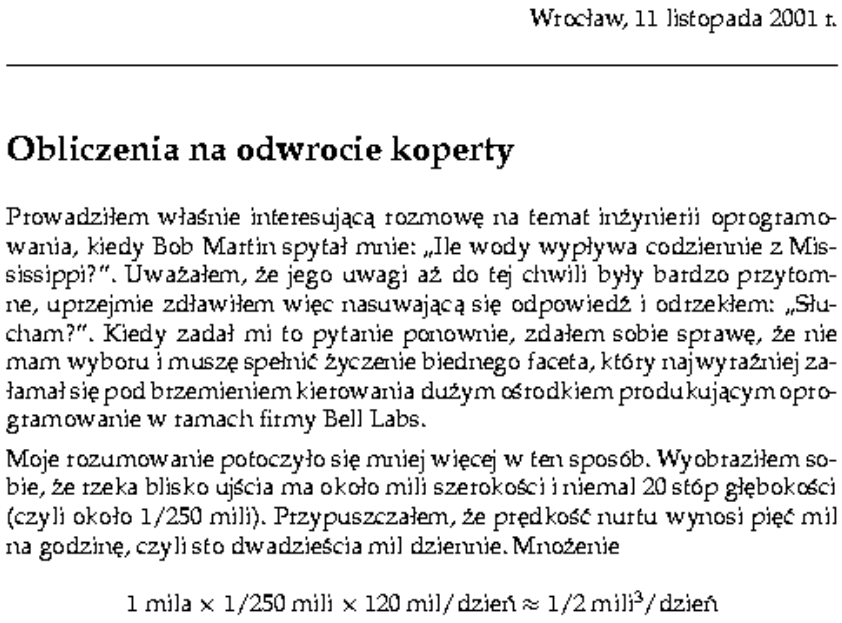
tuł, wprowadzenie, śródtytuł, cytat, akapit), zaś pojęcie sposobu prezentacji odnosi się do określenia wyglądu tych fragmentów (np. decyzje o kroju i wielkości pisma lub szerokości tekstu i głębokości wcięć) [17].
Na rysunku 2.1 przedstawiono przykład pliku tekstowego oraz tę samą treść z oznaczonymi elementami struktury przy pomocy języka LaTeX. Obecność znaczników umożliwia dynamiczne formatowanie dokumentu przy pomocy reguł stylistycznych.

Rysunek 2.1: Przykład pliku: (a) tekstowego, (b) z oznaczoną strukturą za pomocą języka LaTeX. Źródło:[17].

Efekt końcowy widoczny jest na rysunku 2.2. Można w tym miejscu dodać, że niniejsza praca również została przygotowana przy użyciu środowiska LaTeX.

2.2 Struktura logiczna i fizyczna dokumentu

Jak wspomniano w rozdziale 2.1, zazwyczaj w pliku oprócz samej treści umieszcza się informacje o strukturze. Co do zasady na każdy dokument składa się treść (ciąg znaków połączonych strukturą logiczną) i forma, co przedstawia również rysunek 2.3. Forma (układ typograficzny) dokumentu, podlegająca zasadom typografii, jest ostatecznie realizowana przez drukarnię zgodnie ze wskazówkami grafików. Najczęściej wygląd tekstu oryginalnego znacznie różni się od tekstu ostatecznego, złożonego przez drukarnię.



Wrocław, 11 listopada 2001 r.

Obliczenia na odwrocie koperty

Prowadziłem właśnie interesującą rozmowę na temat inżynierii oprogramowania, kiedy Bob Martin spytał mnie: „Ile wody wypływa codziennie z Mississippi?". Uważałem, że jego uwagi aż do tej chwili były bardzo przytomne, uprzejmie zdławiłem więc nasuwającą się odpowiedź i odrzekłem: „Słucham?". Kiedy zadał mi to pytanie ponownie, zdałem sobie sprawę, że nie mam wyboru i muszę spełnić życzenie biednego faceta, który najwyraźniej załamał się pod brzemieniem kierowania dużym ośrodkiem produkującym oprogramowanie w ramach firmy Bell Labs.

Moje rozumowanie potoczyło się mniej więcej w ten sposób. Wyobraziłem sobie, że rzeka blisko ujścia ma około mili szerokości i niemal 20 stóp głębokości (czyli około 1/250 mili). Przypuszczałem, że prędkość nurtu wynosi pięć mil na godzinę, czyli sto dwadzieścia mil dziennie. Mnożenie

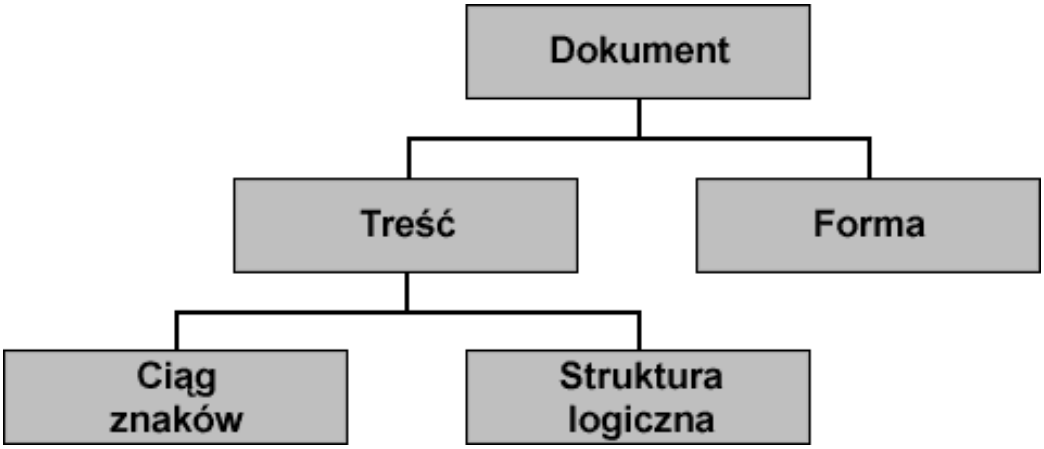
$$1 \text{ mila} \times 1/250 \text{ mili} \times 120 \text{ mil/dzień} \approx 1/2 \text{ mili}^3/\text{dzień}$$

Rysunek 2.2: Wynik formatowania pliku tekstowego. Źródło:[17].

Grafik czy też operator składu, dla przygotowania wydruku, który będzie zgodny z zamysłem autora artykułu, musi prawidłowo zinterpretować znaczenie poszczególnych elementów tekstu i określić ich miejsce w hierarchii dokumentu. Uniknięcie przekłamań i błędów w procesie składu i formatowania jest możliwe tylko wtedy, gdy autor oraz grafik posługują się tym samym kodem semantycznym.

W tym celu konieczne jest stworzenie formalnej listy elementów, których wystąpienie jest obligatoryjne oraz takich, które są opcjonalne, a obecność jest uzależniona od zamysłu autora. Kolejność wystąpienia poszczególnych elementów oraz ich relacje nie pozostają również przypadkowe, lecz są odpowiednio sklasyfikowane. W przypadku większych form pisarskich może ponadto występować kilka alternatywnych szablonów, a decyzja o wyborze ostatecznym może być odłożona w czasie.

Elementy dokumentu i relacje, które je łączą, opisuje się przy pomocy drzewa. Przykład takiego drzewa zaprezentowano na rysunku 2.4.



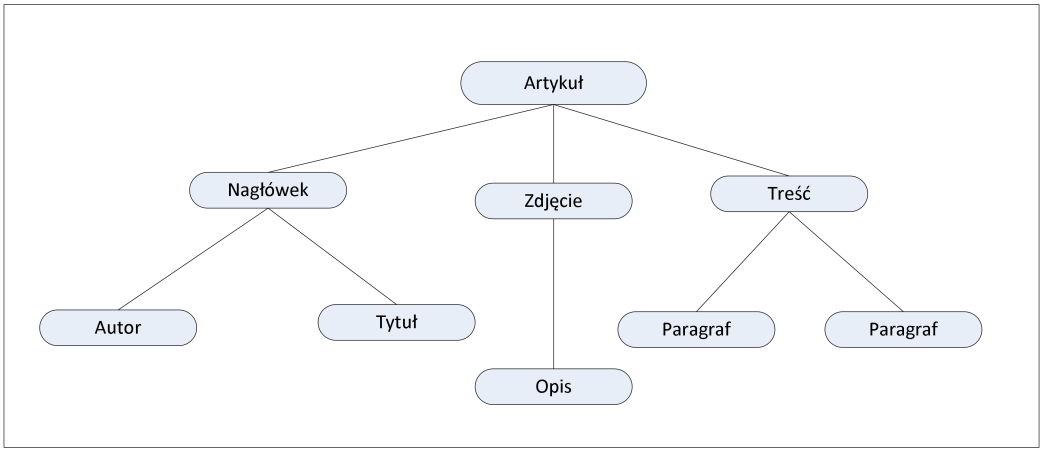
Rysunek 2.3: Ogólna struktura dokumentu.

Na uwagę zasługuje również fakt, że nie wszystkie informacje, które zostały przekazane do redakcji, muszą podlegać publikacji. Artykuł źródłowy zawierać może szereg informacji porządkowych, referencyjnych czy poufnych. Niektóre mogą być wykorzystywane do przygotowania wydań elektronicznych - na przykład aktywne odnośniki do zasobów Internetu, które w wydaniu papierowym nie mają racji bytu.
Chociaż jest technicznie możliwe zapisywanie w pliku PDF pełnej informacji o strukturze dokumentu (poprzez tzw. otagowanie [15]) na chwilę obecną wydawcy operujący na polskim rynku usług monitoringu mediów nie wykorzystują tego faktu.

Na rysunku 2.5 zamieszczono artykuł prasowy, w którym wyodrębniono:

- tytuł,
- nagłówek,
- treść,
- śródtytuł,
- elementy graficzne (zdjęcia, infografika).

Możemy jedynie się domyślać w jaki sposób informacje strukturalne zostały zakodowane przez autora. Nie dysponujemy nimi w sposób bezpośredni,




```
<artykuł>
        <nagłówek>
                <autor>
                        Małgorzata Grzegorczyk
                </autor>
                <tytuł>
                        Bain w Warszawie
                </tytuł>
        </nagłówek>
        <zdjęcie>
                ....
        </zdjęcie>
        <treść>
                <paragraf>
                        ...
                </paragraf>
                <paragraf>
                        ...
                </paragraf>
        </treść>
</artykuł>
```

Rysunek 2.4: Drzewo opisujące strukturę logiczną artykułu oraz reprezentacja w postaci dokumentu XML.

Bain w Warszawie

Biurem amerykańskiej firmy doradczej, które z Polski obsłuży cały region, pokieruje Jacek Poświata

Małgorzata Grzegorczyk
m.grzegorczyk@pb.pl ☎ 22-333-98-56

Amerykańska firma doradztwa strategicznego Bain & Company już kiedyś otworzyła biuro w Warszawie. Ale szybko je zamknęła. Teraz wraca.

– To odpowiedź na stabilny wzrost gospodarczy Polski oraz wzrost znaczenia regionu na świecie – mówi Jacek Poświata, dyrektor zarządzający Bain & Company na Polskę i Europę Środkową i Wschodnią.

Kiedy pół roku temu firma podała na swojej stronie, że myśli o otwarciu biura w Europie Środkowej, do oddziału w Niemczech napłynęło ponad 200 aplikacji, w tym większość z Polski.

Rekrutacja bez anonsu

Kandydaci do pracy słusznie przewidzieli, że regionalne biuro znajdzie się nad Wisłą. Tymczasowa lokalizacja znajduje się w budynku Skylight w Warszawie. W sierpniu firma planuje wprowadzić się do nowego biura w centrum o powierzchni 300-400 mkw. Ponieważ aplikacje już napłynęły, rekrutacja trwa, choć Bain nie ogłosił, że szuka pracowników.

– Chcemy osiągnąć tzw. masę krytyczną, czyli zatrudniać 30 osób w biurze w Warszawie. Część pracowników przeniesiemy z innych oddziałów Bain. Zatrudniamy osoby z doświadczeniem w konsultingu, finansach lub przemyśle, chcemy też dać pracę najzdolniejszym absolwentom – zapowiada Jacek Poświata.

Lubi budować

Zarządzający Bain w regionie, który jest także członkiem rady nadzorczej KGHM, to postać świetnie znana w konsultingu. Rok temu, po 20 latach pracy w firmie, odszedł ze stanowiska dyrektora zarządzającego McKinsey. Doprowadził do rozkwitu spółki w Polsce. McKinsey zatrudnia dziś 40 konsultantów. W ubiegłym roku firma otworzyła centrum wiedzy dla regionu EMEA we Wrocławiu (80 pracowników) i centrum usług wsparcia w Poznaniu (130 osób).

– Od lat doradzam firmom w budowaniu silnej pozycji na rynku, a jednocześnie zawsze fascynowało mnie budowanie biznesu od podstaw. W Bain mogę połączyć jedno z drugim – mówi Jacek Poświata.

Przejście znanego menedżera do konkurencji to na rynku usług doradczych nic nowego.

– Jeśli ktoś z 20-letnim doświadczeniem w branży decyduje się rozwijać nową firmę, musi wiedzieć, że to wypali. Nie oznacza to odebrania klientów, bo o nich walczy się często w przetargach – uważa Michał Gwizda, współwłaściciel Accreo, firmy doradczej założonej przez byłych pracowników Ernst & Young.

Co ich wyróżnia

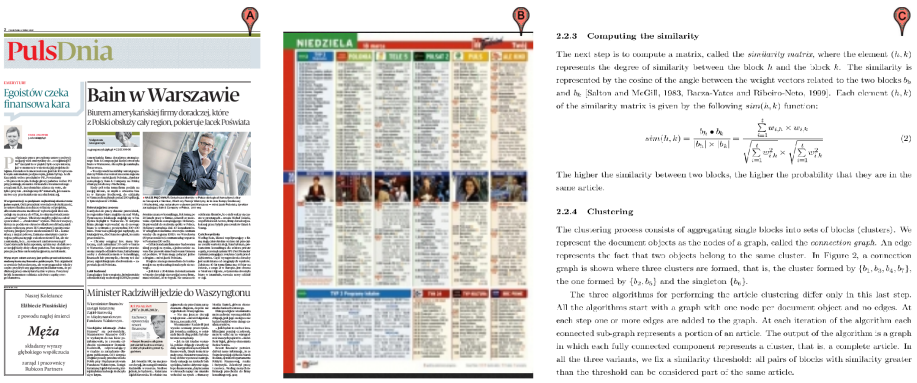
Według Bain, klienci współpracujący z firmą mają czterokrotnie wyższe niż przeciętne zwyżki wartości akcji. Bain był m.in. prekursorem konsultingu dla sektora private equity, rozwinął system NPS (Net Promoter System) pomagający zwiększać zyski przedsiębiorstwa. Część wynagrodzenia doradcy jest uzależniana od osiągniętych wyników. Założona 40 lat temu firma ma 48 biur na świecie, z czego 21 w Europie. Jest obecna w Moskwie i Kijowie, w tym roku otworzyła biuro w Istambule, rozważa nowy oddział w Azji.

NASZE PIĘĆ MINUT: Dotychczas klientów w Polsce obsługiwali konsultanci z biur ze Szwajcarii, z Niemiec, Włoch czy Francji. Wierzymy, że to czas Europy Środkowej i Wschodniej, więc naturalnym wyborem jest Warszawa — mówi Jacek Poświata, dyrektor zarządzający Bain & Company w Polsce. [FOT. WM]

Rysunek 2.5: Przykładowy artykuł prasowy: (a) tytuł, (b) nagłówek, (c) inf. o autorze, (d) treść główna, (e) śródtytuł, (f) grafika, (g) opis grafiki, (h) elementy dekorujące. Źródło: *Puls Biznesu* wydanie z 4 czerwca 2012 r.

niemniej człowiek jest w stanie niemal bezbłędnie odtworzyć strukturę logiczną dokumentu. Podczas czytania gazety czy nawet przelotnego przeglądania kolejnych stron, mózg w sposób bezwiedny dokonuje klasyfikacji treści i przypisuje obszarom odpowiednie role - zazwyczaj bez ich formalnego nazywania. Uwaga czytającego skupia się najpierw na elementach najlepiej widocznych, najbardziej przyciągających wzrok. Może to być wielokolorowe zdjęcie, grafika lub tytuł napisany wielką czcionką - niekiedy jest to zbieg marketingowy, obliczony wyłącznie na zwrócenie uwagi.

Wspomniana klasyfikacja treści jest możliwa z uwagi na fakt, że każdy dokument jest reprezentantem określonej klasy dokumentów i jako jej członek podlega jej regułom. Klasyfikacja grupy dokumentów również jest dla człowieka procesem intuicyjnym i zazwyczaj mimowolnym. Jeden rzut oka wystarcza, aby stwierdzić czy patrzymy na program telewizyjny, gazetę codzienną czy publikację naukową, co zaprezentowano na rysunku 2.6.



Rysunek 2.6: Przykłady różnych klas dokumentów: (a) gazeta codzienna, (b) przewodnik telewizyjny, (c) publikacja naukowa.

Podczas składu gazety, tj. wytworzenia struktury fizycznej dokumentu, przetwarza się jej struktury logiczne i do każdej z nich stosuje się reguły formatowania. Reguły formatowania mają tutaj kluczowe znaczenie, albowiem są one współdzielone przez klasy dokumentów. Jest ogólnie przyjętą normą, że tytuł umieszczany jest (zazwyczaj) nad ciałem artykułu i pisany jest czcionką o największym stopniu, następujące po nim wprowadzenie w postaci 2-5 zdań nieznacznie wyróżnia się od ciała artykułu i streszcza je.

Główna cześć tekstu ujęta jest w bloki zwane akapitami, które rozpoczynają się wcięciem w pierwszej linii oraz oddzielone są między sobą większym odstępem niż poszczególnie linie tekstu w tychże, cytaty pisane są kursywą lub w apostrofach itd.

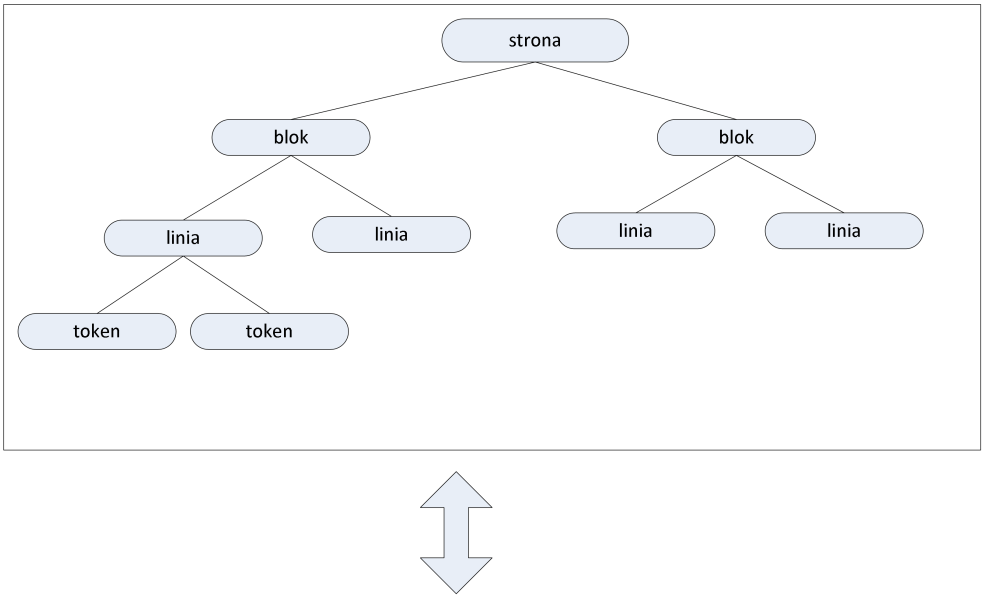
Rysunek 2.7: Fizyczna budowa dokumentu. Wyróżniono elementy z kolejnych poziomów hierarchii.

Na rysunku 2.8 zaprezentowano fragment wcześniej analizowanego artykułu z uwzględnieniem elementów struktury fizycznej. Drzewa struktury logicznej i fizycznej korespondują ze sobą w sposób naturalny. Elementy struktury logicznej, chociaż nieopisane tutaj w sposób bezpośredni, wyrażone są atrybutami typograficznymi.

2.3 Skład tekstu

Za *skład* tekstu odpowiedzialny jest grafik lub operator DTP, który korzysta z wyspecjalizowanego oprogramowania. Czołowe pozycje z tej listy to komercyjne Adobe InDesign[2] czy QuarkXPress[25], ale funkcjonują też ich darmowe odpowiedniki jak Scribus[26] lub PagePlus[23]. Wybór narzędzi darmowych nie wpływa bezpośrednio na jakość końcową produktu, więc przy ich użyciu możliwe jest osiągnięcie tak samo dobrych efektów, niemniej wymagać to może większych nakładów pracy od operatora.
Na dane wejściowe składają się:

- kolekcja dokumentów o charakterze jednostkowym,
- kolekcja artykułów o charakterze ciągłym (np. powieść w odcinkach),
- ogólny szablon publikacji,



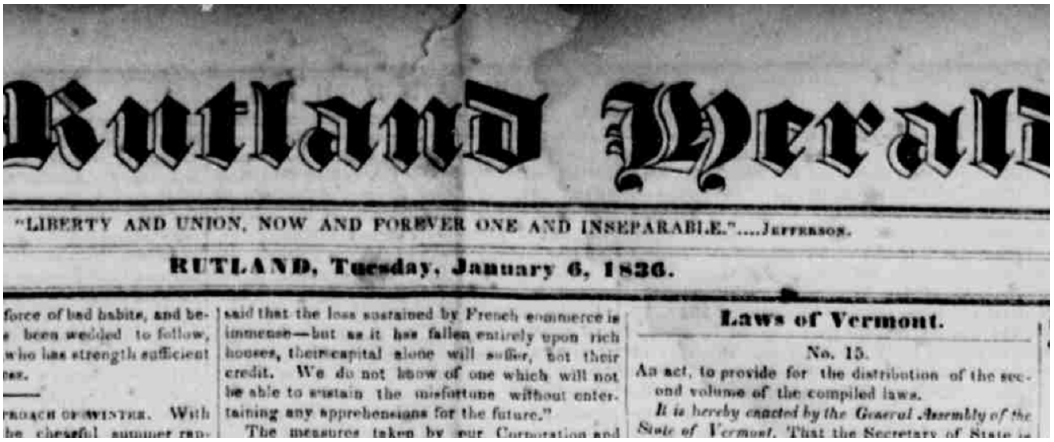

```
<strona>
        <blok>
                <linia>
                        <token czcionka="Calibri">
                                Ba
                        </token>
                        <token czcionka="Calibri">
                                in w War
                        </token>
                        ...
                </linia>
        </blok>
        <blok>
        ...
        </blok
</strona>
```

Rysunek 2.8: Drzewo struktury fizycznej artykułu oraz reprezentacja w postaci dokumentu XML.

- szczególne wytyczne dotyczące formatowania (np. zubożona kolorystyka w dniach żałoby narodowej),

- kolekcja elementów graficznych (np. treści reklamowe, fotoreportaże).

Wyżej wymienione treści są rozkładane zgodnie z zamysłem operatora DTP na poszczególnych stronach publikacji. Nie dysponuje on naturalnie pełną dowolnością, lecz musi zapewnić realizacje szeregu warunków o charakterze obligatoryjnym (rozmiar publikacji, zawarte kontrakty na umieszczenie treści reklamowych, dopasowanie charakteru treści do miejsca publikacji itp.) Jego zadaniem jest również dopasowanie wielkości poszczególnych elementów (np. wyrównanie wielkości stopnia pisma dla tytułów i śródtytułów), dobór fontów, korekcja obrazów rastrowych. Poszczególne technologie samego druku posiadają również swoje odrębne ograniczenia, które wynikają ze stosowanych technik czy komponentów (ograniczona paleta barw, mniejsza rozdzielczość). Kryterium oceny jego pracy jest *jak najlepszy efekt wizualny* uzyskany po wydruku; jest to kryterium subiektywne. Gotowy projekt jest przesyłany do drukarni w postaci pliku PDF (jednego lub w rozbiciu na kolory bazowe np. CMYK), a także może również być publikowany w kanałach dystrybucji elektronicznej wydawnictwa. Tak wytworzony plik stanowi również punkt wyjściowy do prac firm pressclippingowych.


Rutland Herald

"LIBERTY AND UNION, NOW AND FOREVER ONE AND INSEPARABLE."....JEFFERSON.

RUTLAND, Tuesday, January 6, 1836.

force of bad habits, and be- been wedded to follow, who has strength sufficient ter.

said that the loss sustained by French commerce is immense—but as it has fallen entirely upon rich houses, their capital alone will suffer, not their credit. We do not know of one which will not be able to sustain the misfortune without entertaining any apprehensions for the future."

Laws of Vermont.

No. 15.

An act, to provide for the distribution of the second volume of the compiled laws.

It is hereby enacted by the General Assembly of the State of Vermont, That the Secretary of State is

Rysunek 2.9: Skan dokumentu z 1836 r. Źródło:[19].

W ogólnym przypadku przyjąć należy, że dokument mógł powstać w dowolny inny sposób. Szczególnie w ostatnich latach upowszechnił się proces digitalizacji dokumentów papierowych, tak w przypadku firm komercyjnych, jak i wszelakich instytucji państwowych oraz organizacji pożytku publicznego. Biblioteki i archiwa, które nierzadko dysponują dokumentami czy publikacjami sprzed stu i więcej lat (patrz rysunek 2.9), są szczególnie zainteresowane utrwaleniem swoich zasobów w postaci cyfrowej. Z jednej strony umożliwia to udostępnianie materiałów szerokiej masie podmiotów, z drugiej - poprzez analizę i indeksację - pozwala na budowanie pełnotekstowych baz danych.

Dokumenty PDF, które są jedynie kontenerem dla obrazu rastrowego, nie mogą być przetwarzane przy pomocy opisywanych algorytmów w sposób bezpośredni. Dla ekstrakcji tekstu konieczne jest przynajmniej przeprowadzenie procesu rozpoznania przy pomocy technik OCR.

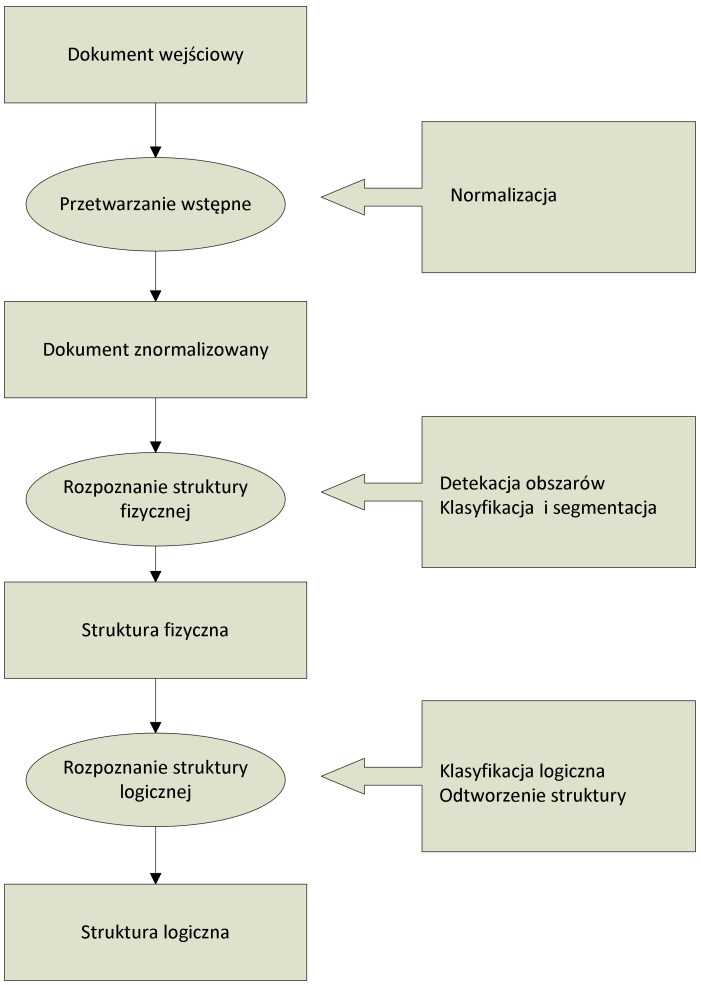
2.4 Analiza dokumentu

Procesem odwrotnym do procesu generowania dokumentu jest jego *analiza i rozpoznawanie*. Jest on stosowany dla wszystkich rodzajów dokumentów, zarówno skanowanych (rastrowych), jak i elektronicznych (PDF). W procesie tym podejmuje się próbę odzyskania pewnych informacji, które podczas wytwarzania mogły zostać utracone lub zniekształcone.
Składa się on z kilku etapów:

- normalizacja danych wejściowych,
- rozpoznanie struktury fizycznej,
- odtworzenie struktury logicznej,
- zapis danych wyjściowych,

co pokazano na rysunku 2.10.

Wymienione fazy mogą - w zależności od zastosowanego algorytmu i rodzaju danych wejściowych - występować sekwencyjnie lub iteracyjnie.



Rysunek 2.10: Proces wieloetapowej analizy dokumentu. Na podstawie:[7].

2.4.1 Przetwarzanie wstępne

Gdy dane wejściowe dostarczone są w postaci zadrukowanego papieru, pierwszym krokiem jest jego digitalizacja w celu otrzymania elektronicznej wersji dokumentu np. w postaci obrazu rastrowego. Tak otrzymany obraz dokumentu może zawierać szumy lub inne artefakty, które dołączyły się na etapie skanowania i mogą być spowodowane przez kurz, nieodpowiednie parametry skanowania, uszkodzenia papieru.
Istotą przetwarzania dokumentu na tym etapie jest normalizacja danych wejściowych w celu zwiększenia skuteczności działania w dalszych krokach procesu. Normalizacja może obejmować zmiany wynikające ze:

- specyfiki algorytmów zastosowanych w dalszych krokach (np. zmiana układu współrzędnych, kodowania),

- specyfiki oczekiwanych danych wyjściowych (np. konwersja przestrzeni barw).

W tej fazie stosuje się takie operacje jak usuwanie szumów, detekcję i korekcję pochylenia, usuwanie tła, wygładzanie, progowanie i tym podobne.

2.4.2 Analiza struktury fizycznej

W procesie analizy i budowy drzewa struktury fizycznej dokumentu głównym celem jest dekompozycja i stworzenie hierarchii regionów homogenicznych takich jak: obrazy, grafiki czy bloki tekstu. Metody analizy fizycznej, w zależności od przyjętej strategii działania, dzielą się na trzy główne kategorie[3, 9, 11]:

- *wstępujące (bottom-up)* - algorytm startuje od najmniejszych komponentów dokumentu i w sposób iteracyjny łączy je w coraz większe homogeniczne regiony np. litery łączy w słowa, słowa w linie, linie w akapity itd. Główną zaletą tej metody jest fakt, że może ona poradzić sobie z dowolnie ukształtowanym obszarem.

- *zstępujące (top-down)* - algorytm w chwili startu analizuje dokument jako całość, na najwyższym poziomie hierarchii, a następnie rekurencyjnie dzieli go na obszary homogeniczne przy pomocy zdefiniowanych reguł detekcji. Metoda ta dobrze sprawdza się przy dokumentach o bardzo prostym układzie (jak chociażby niniejsza praca, w której tekst umieszczony jest w jednej, ciągłej kolumnie i czasem przedzielony jest rysunkami czy tabelami). W przypadkach bardziej skomplikowanych - co ma miejsce w przypadkach e-wydań prasowych - skuteczność tej metody jest znacznie niższa, ze względu na trudność wyznaczenia generycznych kryteriów podziału.

- *mieszane* - jest kombinacją dwóch wyżej wymienionych strategii

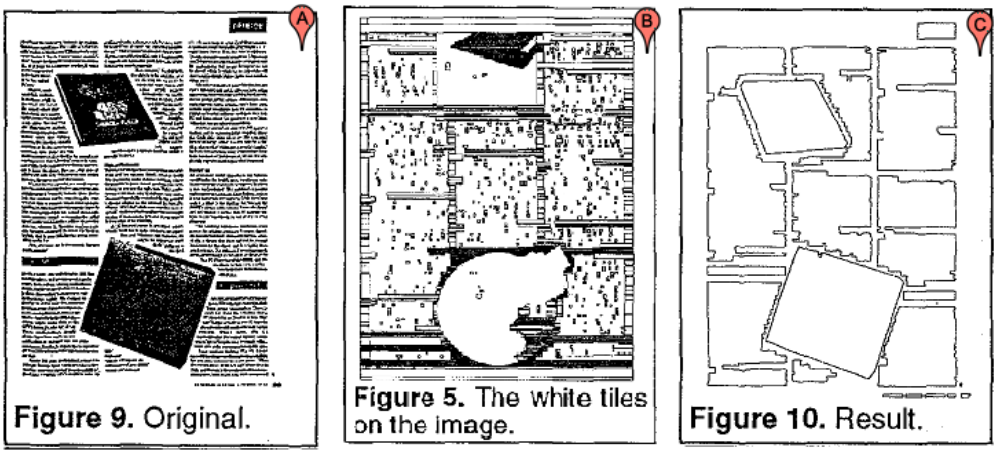
Białe kafle

W artykułach [3, 4] zaproponowano metodę segmentacji, która działa w oparciu o obraz rastrowy. Przyjęto, że dokument składa się regionów tekstowych przeplatanych elementami graficznymi, przy czym każdy z obszarów może być dowolnego kształtu. Dookoła każdego obszaru znajdują się białe miejsca, które go ograniczają. W kontekście całej strony, wszystkie obszary łączą się w sieć, w której oka są różnych kształtów i rozmiarów i odpowiadają obszarom zadrukowanym, co prezentuje rysunek 2.11

O ile obszar jest prostokątem obejmującym wiele linii tekstu, może zostać podzielony na wiele podobszarów o wysokości linii tekstu, z którą zostaje powiązany. W procesie segmentacji, kontury obszarów zadrukowanych są wyznaczane poprzez analizę granic kafli. Na podstawie wcięcia w pierwszej linii tekstu, wykrywane są akapity. W artykule nadmieniono, że algorytm znajduje również zastosowanie dla dokumentów, które mają pionową orientację druku (np. pismo japońskie), a jako główne zalety wymieniono szybkość i dokładność działania.

Segmentacja tekstu

Odmienne założenia zostały przyjęte w artykule [9]. Przyjęto, że algorytm dedykowany jest dla *prawdziwych* dokumentów PDF. Nie obejmuje więc ta-



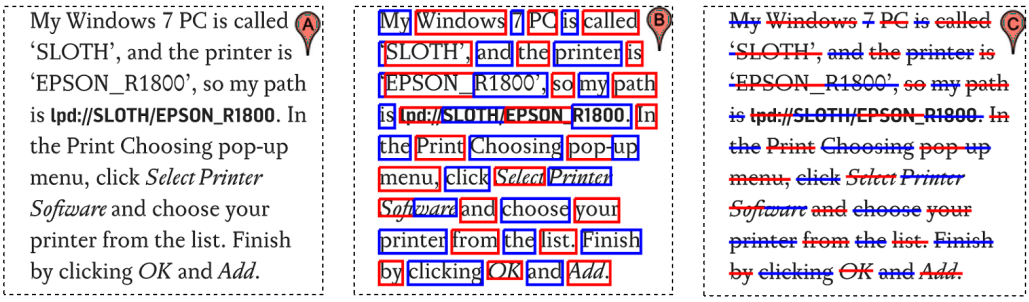
Rysunek 2.11: Przykład działania algorytmu *white tiles*: (a) obraz źródłowy, (b) wyodrębnione obszary bez zadruku, (c) granice obszarów tekstowych. Źródło:[3].

kich, które zawierają jedynie opakowany obraz rastrowy. Przyjęto, że tekst może być wyselekcjonowany i oddzielony od obiektów graficznych i obrazów przy pomocy dostępnych bibliotek. Porządek czytania tekstu powinien być z góry na dół, przy czym tekst ułożony jest w poziome linie. Algorytm składa się z trzech głównych faz:

1. pozyskanie informacji o obiektach tekstowych
2. łączenie słów w linie
3. grupowanie linii tekstu w bloki

W fazie pierwszej posłużono się biblioteką pozyskaną z firmy Datalogics. Za pomocą komponentu WordFinder dokonano ekstrakcji słów wraz z ich atrybutami wizualnymi takimi jak: rodzaj czcionki, jej wielkość, kolor oraz prostokąt otaczający. Z niniejszego artykułu zaczerpnięto przykład pokazany na rysunku 2.12. Uwagę zwraca fakt, że wysokość obszaru otaczającego (rys. 2.12b) jest zmienna w kolejnych liniach, a nawet obserwujemy fluktuacje w obrębie jednej linii z uwagi na zmiany wielkości czcionki. Mniejsze

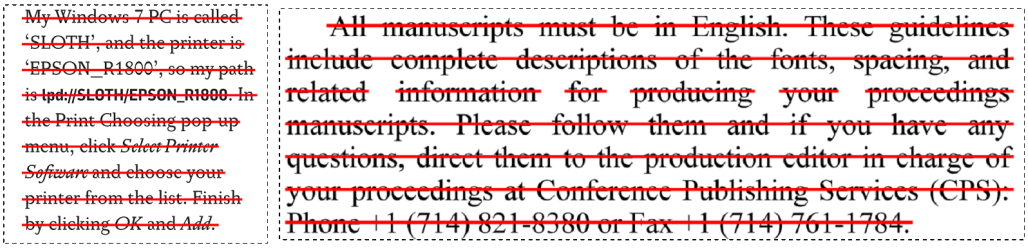
wahania obserwujemy dla linii wyznaczających środki symetrii otaczających prostokątów (rys. 2.12c).



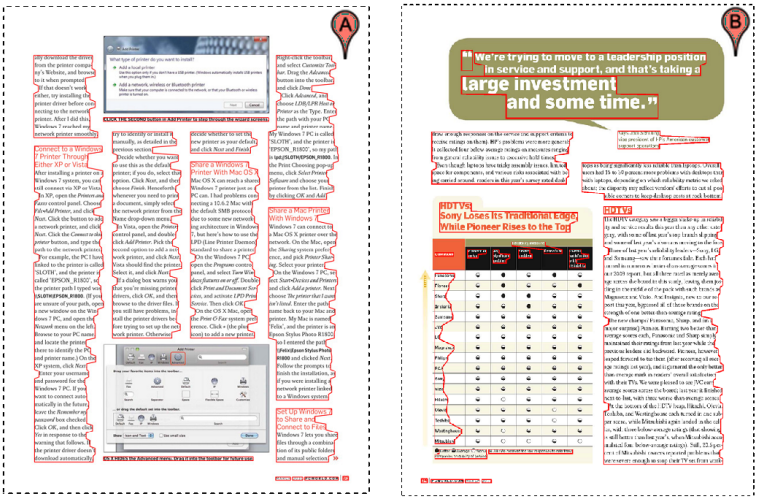
Rysunek 2.12: Wyodrębnienie słów w tekście: (a) przykładowy tekst, (b) wyodrębnione słowa otoczone prostokątem, (c) poziome linie wyznaczają środek otoczenia słowa. Źródło:[9].

W kroku drugim wyodrębnione wcześniej słowa łączone są w linie. Ponieważ kolejność renderowania obiektów nie jest tożsama z kolejnością czytania, dalsze przetwarzanie opiera się na atrybutach przestrzennych oraz wielkości czcionki.
Proces formowania linii rozpoczyna się od posortowania kolekcji obszarów w porządku góra-dół, lewo-prawo bazując na punkcie środkowym prostokąta otaczającego. Następnie pobierane są kolejne nieprzetworzone elementy i na podstawie kryteriów nachodzenia lub wzajemnego odstępu obszarów oraz zmiany wielkości czcionki podejmowana jest decyzja o przyporządkowaniu obszaru do konkretnej linii. Przykładowe wyniki pokazano na rysunku 2.13.

W kroku trzecim linie tekstu łączone są w większe bloki począwszy od organizacji fragmentów linii w jedną logiczną linię, aż po łączenie linii w akapity. Etap ten jest szczególnie wrażliwy na dobór algorytmicznych współczynników liczbowych, określających próg przełączania kwalifikatora. Eksperymentalnie dowiedziono, że niemożliwe jest określenie arbitralnych wartości, które byłyby poprawne dla wszystkich przypadków. Dobór tych wartości jest często kwestią indywidualną dla dokumentu.



Rysunek 2.13: Dane po drugiej fazie przetwarzania - słowa połączone w linie. Źródło:[9].



Rysunek 2.14: Dane wynikowe: (a) poprawnie rozpoznane granice tekstu, (b) błędna klasyfikacja elementów tabeli. Źródło:[9].

Na rysunku 2.14 pokazano wyniki końcowe działania algorytmu. Rysunek (a) prezentuje poprawne wyniki, natomiast na rysunku (b) widzimy przykład strony, dla której algorytm ten nie był dość skuteczny - tabela nie została dobrze rozpoznana. Takie obiekty jak tabele, mapy czy wykazy często wymagają przetwarzania przy użyciu algorytmów bardziej specjalizowanych [13].

Zintegrowany algorytm dekompozycji

Ostatni z omawianych algorytmów został przedstawiony w artykule [11]. Z dokumentu wejściowego kolejno dokonywana jest ekstrakcja linii, obrazów i elementów graficznych oraz bloków tekstowych. Dla każdego z tych procesów zaproponowano w artykule użycie nowych metod i metryk klasyfikacji. Ekstrakcja linii oparta jest na przekształceniu Hougha oraz operacjach morfologicznych. Uznano, że informacja o obecnych w dokumencie liniach będzie szczególnie przydatna w procesie segmentacji artykułów, co zaprezentowano na rysunku 2.15.

Rysunek 2.15: Ekstrakcja linii z dokumentu. Źródło:[11]

Ekstrakcję obrazów i elementów graficznych, prowadzi się w oparciu o trans-

formatę FFT. Algorytm umożliwia również poprawną ekstrakcję opisów obrazów, które w tej sytuacji powinny być potraktowane jako pełnoprawny obiekt tekstowy, a nie jako część obrazu. Pełny wynik działania zaprezentowano na rysunku 2.16.

Rysunek 2.16: Dekompozycja strony gazety (a) strona oryginalna, (b) obraz po segmentacji, (c) zidentyfikowane artykuły. Źródło:[11]

2.4.3 Analiza struktury logicznej

Rozpoznawanie struktury logicznej jest procesem odwrotnym do procesu formatowania dokumentu podczas jego tworzenia. W tym procesie, na podstawie struktury fizycznej dokumentu, wnioskuje się o strukturze logicznej. W przypadku artykułów prasowych tytuł jest pisany większą czcionką, wprowadzenie następuje po tytule, śródtytuł jest umieszczony między akapitami i może być wytłuszczony itd.

Analiza struktury logicznej ma na celu wyodrębnienie z dokumentu elementów logicznych oraz ich wzajemnych relacji. Bazuje ona na wynikach analizy fizycznej i obejmuje dwie fazy:

- *etykietowanie* - określenie logicznej funkcji obszaru i jego atrybutów,
- *rekonstrukcję hierarchii* - połączenie elementów w drzewo struktury logicznej.

W procesie analizy logicznej uwaga skupiona jest również na pojęciu porządku czytania, tj. kolejności wystąpienia bloków tekstu w sekwencji. Takie języki jak arabski, chiński czy japoński mogą mieć porządek odmienny (prawy do lewej, z góry do dołu) niż ma to miejsce w językach łacińskich. Z tego powodu, uniwersalny algorytm wielojęzykowy powinien uwzględniać i parametryzować tę kwestię.

Większość obecnie powstających algorytmów i narzędzi służących do analizy dokumentów PDF ma charakter kompleksowy, tzn. obejmuje zarówno etap analizy fizycznej jak i logicznej. Dotyczy to również algorytmów opisanych w rozdziale 2.4.2. W niniejszym rozdziale przedstawione zostaną te, które na analizę logiczną i identyfikację artykułów kładą szczególny nacisk.

Algorytm grafowy

Algorytm grafowy został opisany w artykule [6]. Autorzy skupili się na problemie identyfikacji i wyodrębnienia artykułów ze strony. Zaproponowali rozwiązanie bazującej na teorii grafów. Przyjęli, że każdy wyodrębniony na stronie element (obszar) jest węzłem grafu, oraz węzły są ze sobą połączone, o ile elementy przynależą do tego samego artykułu. Przynależność do artykułu (a więc decyzja o utworzeniu połączenia pomiędzy węzłami) jest podejmowana poprzez obliczenie wartości funkcji metrykalnej i porównaniu jej z wartością progową.

Rysunek 2.17: Przykład analizy logicznej z użyciem grafu: (a) strona oryginalna, (b) wydzielone bloki artykułów wraz z oznaczeniem numerycznym, (c) reprezentacja grafowa - każdy blok reprezentowany jest przez węzeł grafu i połączony jest z innymi węzłami przynależnymi do artykułu. Źródło:[6].

W artykule zaproponowano ponadto użycie dwóch metryk dla oceny ja-

kości działania algorytmu klasteryzacji.

Sztuczna inteligencja

Odmienne podejście zaprezentowano w artykule [8]. Założono, że system analizy powinien być na tyle funkcjonalny, aby zapewnić poprawne wyniki dla różnych rodzajów dokumentów wejściowych - powinien być więc adaptacyjny. Za przykład wzięto *człowieka*, który będąc w nowej, nieznanej sobie sytuacji potrafi podjąć odpowiednie działania wykorzystując dotychczas zebraną wiedzą ogólną. W kontekście przetwarzania dokumentów, taką nieznaną sytuacją będzie pojawienie się nowego typu dokumentu - takiego, z którym wcześniej system nie pracował. Przyjęto, że system - tak jak człowiek - powinien być w stanie przyswoić sobie niezbędną wiedzę tak, aby w dalszym toku mógł poprawnie przetwarzać dokumenty tej klasy.

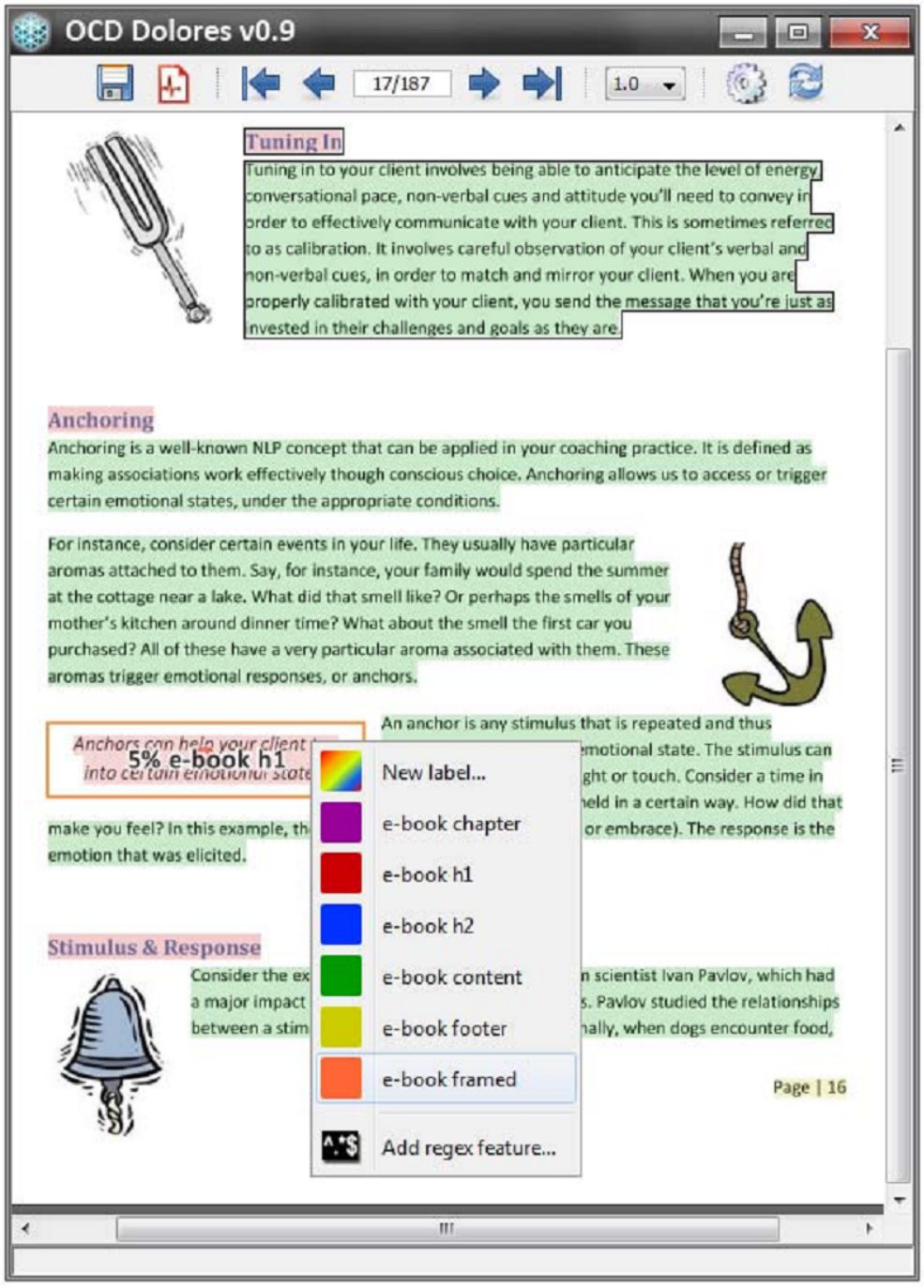
Typowy przebieg interakcji z użytkownikiem przebiega następująco:

- system wczytuje dane i je analizuje,
- system przedstawia użytkownikowi propozycję rozwiązania,
- człowiek sprawdza poprawność i wprowadza potrzebne korekty,
- system dokonuje adaptacji swojego modelu na podstawie wprowadzonych przez człowieka korekt.

System *OCD Dolores* jest zbudowany w oparciu o sztuczną sieć neuronową. W celu zminimalizowania wkładu człowieka, na początkowym etapie, przygotowuje się dane uczące do trenowania sieci.

Na rysunku 2.18 pokazano przykładową sesję systemu Dolores w trybie uczenia. Za pomocą różnokolorowego tła wyróżnione są poszczególne klasy elementów strony. Zadaniem użytkownika jest sprawdzenie poprawności klasyfikacji. W miarę potrzeby, może on też wprowadzić nowy typ elementu.

Sieć neuronowa działa w oparciu o wektor danych wejściowych. W systemie Dolores zdefiniowano kilkadziesiąt różnorodnych cech, których wartości możliwe są do pozyskania w procesie analizy fizycznej oraz etykietowania:

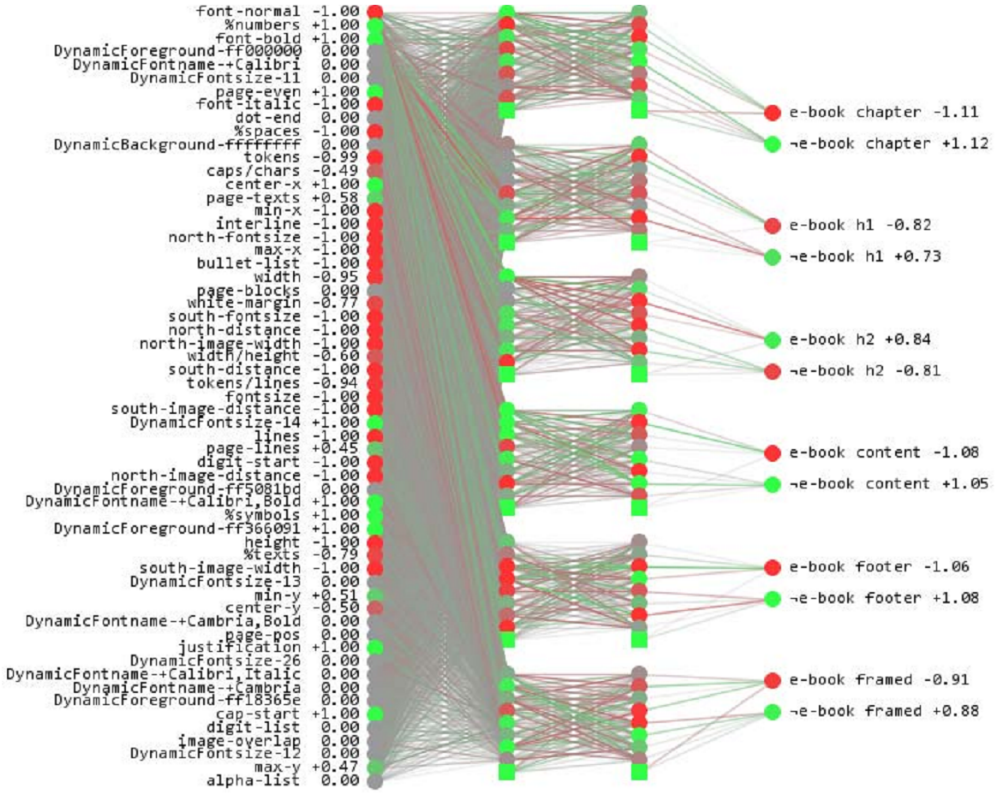


Rysunek 2.18: Interfejs graficzny systemu Dolores w trybie uczenia. Źródło:[8].

- morfologiczne - współrzędne brzegowe i uśrednione, kierunek tekstu rotacja, itp.
- strukturalne - wielkość czcionki, krój czcionki, justowanie, liczba linii, itp.
- syntanktyczne - litery, cyfry, symbole, odstępy, itp.
- relacyjne - szerokość przyległych bloków, rozmiar przyległych bloków, orientacja, itp.
- dynamiczne - generowane dynamicznie w procesie etykietowania

Sieć została zaimplementowana jako wielowarstwowa sieć perceptronowa z własnymi udoskonaleniami autorów. W artykule podkreślono zalety takiego wyboru wskazując na to, że sieć ta wymaga relatywnie niewielkiej liczby danych uczących. Przykładowy układ klasyfikatora rodzaju publikacji zaprezentowano na rysunku 2.19.

Wyniki przeprowadzonych przez autorów badań potwierdzają wysoką skuteczność i możliwości adaptacyjne tego systemu.



Rysunek 2.19: Topologia sieci SMLP z sześcioma neuronami wyjściowymi. Źródło:[8].

Rozdział 3

Elementy przetwarzania obrazów

W niniejszym rozdziale przedstawiono teoretyczne podstawy cyfrowego przetwarzania obrazów. Z uwagi na skomplikowanie procesu tworzenia i formatowania dokumentu, o którym pisano w rozdziale 2.1, specyfikę formatu PDF i brak możliwości przechowania informacji o strukturze dokumentu logicznego, o czym pisano w rozdziale 1, badania literaturowe, a także wyniki wstępnych eksperymentów, zdecydowano się na wsparcie procesu analizy zawartości natywnej poprzez analizę graficzną. Można również zauważyć, że wyłącznie takim sposobem posługuje się podczas czytania człowiek, co świadczy o praktycznej przydatności tych informacji.
Poniżej opisano jedynie te zagadnienia, które znalazły praktyczne zastosowanie w niniejszej pracy.

3.1 Filtracja

W ogólnym procesie analizy zazwyczaj przyjmuje się, że dane wejściowe spełniają określone kryteria jakości (stopień zaszumienia, poziom kontrastu, itp.). Jeżeli obraz wejściowy odbiega o przyjętych czy wymaganych form, konieczne jest jego dostosowanie. Często proces ten nazywany jest filtracją, przy czym w zależności od rodzaju danych wejściowych oraz specyfiki zagad-

nienia, może sprowadzać się do prostego wygładzania konturów i usuwania szumu lub - w bardziej zaawansowanych przypadkach - oznaczać selektywną ekstrakcję określonych fragmentów obrazu lub ich sekwencji.

W filtracji rozumianej jako proces usuwania szumu, zazwyczaj wykorzystuje się informacje o bezpośrednim sąsiedztwie każdego piksela. Poprzez takie niewielkie uogólnienie można z pewnym prawdopodobieństwem wnioskować o tym, czy badany piksel jest pożądany, czy też nie.

Obraz przetwarza się przy użyciu tzw. maski, która może mieć dowolny rozmiar oraz kształt. Maska jest kolejno przesuwana po obrazie (tak aby każdy piksel obrazu wejściowego został przetworzony), a wyniki zapisuje się w obrazie wyjściowym.

$$\begin{bmatrix} (x-1,y+1) & (x,y+1) & (x+1,y+1) \\ (x-1,y) & (x,y) & (x+1,y) \\ (x-1,y-1) & (x,y-1) & (x+1,y-1) \end{bmatrix}$$

Zwyczajowo maska jest macierzą o nieparzystej liczbie wierszy i kolumn - tak aby łatwo można było wyznaczyć punkt centralny, w którym umieszcza się badany piksel, ale może też mieć inny kształt np. prostokąta, rombu, krzyża, koła. Należy zauważyć, że zmniejszenie otoczenia wokół badanego piksela jest równoznaczne ze zmniejszeniem ilości informacji wejściowych, wobec czego skuteczność filtracji jest przez to osłabiona. Dla zwiększenia ważności wybranych pikseli, maska może zawierać dowolne współczynniki liczbowe, przez które mnoży się wartość jasności odpowiadającego piksela.

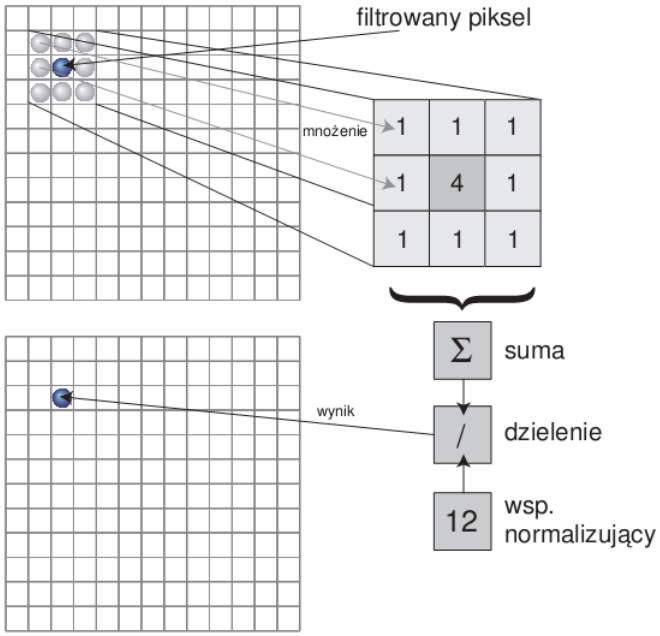
3.1.1 Filtry konwolucyjne

Jedną z najpopularniejszych metod filtracji obrazu jest wykorzystanie operacji konwolucji, zwanej także splotem funkcji [30].Jeżeli funkcje f oraz g są ciągłe i różniczkowalne, to konwolucję definiujemy jako:

$$f \star g = \int f(x-t)g(t)dt$$

W przypadku gdy te funkcje są dyskretne, konwolucję można łatwo opisać

w sposób algorytmiczny i należy tutaj zauważyć, że obraz jest taką funkcją. Dalsze uproszczenie niektórych operacji jest możliwe również dzięki temu, że konwolucja ma właściwości matematyczne zbliżone do operacji mnożenia. W kontekście filtracji obrazów, jako pierwszą z funkcji przyjmuje się obraz wejściowy, natomiast drugą, określa się mianem jądra konwulcji, od postaci którego zależy wygląd obrazu po filtracji.



Rysunek 3.1: Zasada działania dyskretnego splotu. Źródło:[27]

W przypadku splotu dyskretnego maska filtra (jądro) jest kwadratową tablicą zawierającą współczynniki, a zasadę działania zobrazowano na rysunku 3.1.

Maska filtra nakładana jest na obraz w taki sposób, aby jej punkt centralny pokrywał się analizowanym pikselem. Dla wyznaczenia wartości wyjściowej wykonuje się operację mnożenia wartości filtra przez wartość znajdujących się pod nim pikseli (zazwyczaj jest to wartość jasności piksela), a następnie sumuje te wyniki. Z uwagi na fakt, że wynik tej operacji może

wykraczać poza zakres dopuszczalnych wartości, dokonuje się normalizacji. Polega ona na podzieleniu wyniku sumowania przez sumę wszystkich współczynników filtra z zastrzeżeniem, że jeżeli suma współczynników filtra wynosi zero, to współczynnik normalizujący wynosi jeden. Ostateczny wynik zapisywany jest do nowej macierzy obrazu. Próba zapisania do obrazu wejściowego doprowadziłaby do sytuacji, gdzie kolejne piksele byłyby filtrowane z wykorzystaniem pikseli już przefiltrowanych, co dałoby niespodziewane i nieprzewidywalne rezultaty.
Poniżej zaprezentowano wybrane maski filtrów[27]:

Najprostszy filtr rozmywający:

$$\begin{bmatrix} 1 & 1 & 1 \\ 1 & \mathbf{1} & 1 \\ 1 & 1 & 1 \end{bmatrix}$$

Filtr dolnoprzepustowy:

$$\begin{bmatrix} 1 & 1 & 1 \\ 1 & \mathbf{4} & 1 \\ 1 & 1 & 1 \end{bmatrix}$$

Rozmycie gausowskie:

$$\begin{bmatrix} 1 & 2 & 1 \\ 2 & \mathbf{4} & 2 \\ 1 & 2 & 1 \end{bmatrix}$$

Wyostrzenie:

$$\begin{bmatrix} -1 & -1 & -1 \\ -1 & \mathbf{9} & -1 \\ -1 & -1 & -1 \end{bmatrix}$$

3.1.2 Filtry morfologiczne

Prace nad morfologią matematyczną zostały zainicjowane na początku XX wieku przez H. Minkowskiego, lecz początek jej praktycznego wykorzystanie datuje się na lata 60-te. Według koncepcji przedstawionej przez G. Matherno oraz J. Serra bardzo istotne jest przetwarzanie obrazu z wykorzystaniem elementu strukturalnego[21][16]. Element strukturalny jest to pewien wycinek obrazu z wyróżnionym jednym punktem wiodącym. Zazwyczaj jest to koło o jednostkowym promieniu.

Operacje podstawowe

Podstawowymi operacjami morfologicznymi nazywamy takie czynności, których nie da się rozbić na mniejsze jednostki ani zapisać przy pomocy innych operacji. Zalicza się do nich: infimum, supremum, progowanie, dylację, erozję, różnicę[29][16].

Operacja **infimum** dwóch obrazów f_1, f_2 w punkcie (x, y) stanowi ta z funkcji, której wartość dla danego piksela jest mniejsza:

$$(f_1 \cap f_2)(x, y) = min\{f_1(x, y); f_2(x, y)\},$$

gdzie: $f_1(x, y), f_2(x, y)$ - wartości poziomów szarości pikseli dwóch obrazów wejściowych.

Operacja **supremum** dwóch obrazów f_1, f_2 w punkcie (x, y) stanowi ta z funkcji, której wartość dla danego piksela jest większa:

$$(f_1 \cup f_2)(x, y) = max\{f_1(x, y); f_2(x, y)\},$$

gdzie: $f_1(x, y), f_2(x, y)$ - wartości poziomów szarości pikseli dwóch obrazów wejściowych.

Różnica dwóch obrazów jest to takie przekształcenie obrazów opisanymi funkcjami f_1, f_2, gdzie powstaje obraz, dla którego poziom szarości pikseli wyjściowych jest różnicą odpowiadających sobie pikseli z obrazów wejścio-

wych f_1 oraz f_2. Operację tę opisuje się następująco:

$$(f_1 - f_2)(x,y) = \begin{cases} f_1(x,y) - f_2(x,y) & \text{gdy} f_1(x,y) \geqslant f_2(x,y) \\ 0 & \text{w pozostałych wypadkach} \end{cases}$$

Progowanie jest operacją polegającą na zmianie wartości poziomu piksela na jedną z dwóch wartości, w zależności od wyniku porównania odpowiednich pikseli obrazów wejściowych f_1 i f_2. Jedną z nich jest wartość maksymalna, jaką może przyjąć poziom szarości danego piksela, przy czym nie jest istotne ani jaka to jest wartość, ani czy jest ona obecna w obrazie wejściowym. W przypadku niespełnienia zadanego kryterium wyboru, wartość szarości piksela przyjmie drugą z możliwych wartości - wartość zero. Jako kryterium progowania można na przykład przyjąć konkretną wartość poziomu szarości piksela

W wielu przypadkach implementacji operacji morfologicznych stosuje się uproszczenie, polegające na traktowaniu obrazów binarnych jako zbiorów algebraicznych z wyróżnieniem elementów będących tłem (oznaczanych zazwyczaj jako 0) oraz elementów będących obrazem właściwym (oznaczanych zazwyczaj jako 1). Kolejne operacje opisuje się przy założeniu, że obraz został zdefiniowany zgodnie z opisanymi założeniami.

Dylatacja obrazu (zbioru) A przez element strukturalny B jest zdefiniowana jako suma przesunięć zbioru A o wszystkie elementy b, przy czym $b \in B$

$$A \oplus_s B = \bigcup_{b \in B} A_b$$

.

Praktycznie wyznacza się ją przemieszczając element strukturalny B po wszystkich elementach obrazu A. Wartość elementu, dla którego wyznaczana jest dylatacja, jest wówczas sumą logiczną elementu strukturalnego B i części obrazu A, który jest nim przysłonięty. Dylację charakteryzuje ekstensywność, monotoniczność oraz niezmienniczość względem przesunięcia.

Erozja obrazu (zbioru) A przez element strukturalny B jest zdefiniowana

jako przecięcie przesunięć zbioru A o wszystkie elementy b, przy czym $b \in B$

$$A \ominus_s B = \bigcap_{b \in B} A_{-b}$$

.

Podobnie jak w przypadku dylacji, wyznacza się ją przemieszczając element strukturalny B po wszystkich elementach obrazu A. Wartość elementu, dla którego wyznaczana jest erozja, jest wówczas iloczynem logicznym elementu strukturalnego B i części obrazu A, który jest nim przysłonięty. Erozja jest również niezmiennicza i monotoniczna, a także antyekstensywna.

Operacje złożone

Kolejne operacje definiuje się jako złożenie operacji podstawowych, przy czym poprzez ilość tychże złożeń określa się rząd, a tym samym stopień skomplikowania danej operacji[29].

Otwarcie dla obrazów reprezentowanych przez zbiory jest operacją powstającą przez złożenie najpierw operacji erozji, a następnie dylacji:

$$A \circ_s B = [A \ominus_s B] \oplus_s B$$

.

W wyniku działania tej operacji usuwane są z obrazu drobne obiekty i drobne szczegóły. Zasadnicza wielkość dużych figur na obrazie nie ulega zmianie.

Zamknięcie dla obrazów reprezentowanych przez zbiory jest operacją powstającą przez złożenie najpierw operacji dylacji, a następnie erozji:

$$A \bullet_s B = [A \oplus_s B] \ominus_s B$$

.

W wyniku działania tej operacji wypełniane są wąskie wcięcia i zatoki oraz drobne otwory wewnątrz obiektu. Możliwe jest złączenie obiektów leżących blisko siebie, ale zasadnicza wielkość dużych figur również nie ulega zmianie.

Operacja **hit or miss** ("trafi nie trafi") znalazła szczególne znaczenie

w niniejszej pracy. Definiuje się ją przy założeniu, że dane są dwa elementy strukturalne B_1 i B_2 takie, że $B_1 \cap B_2 = \oslash$, a operację określa wyrażeniem:

$$A \otimes (B_1, B_2) = \{h : B_{1h} \subseteq A \text{ oraz } B_{2h} \subseteq A^C\}.$$

Piksel h należy do zbioru $A \otimes (B_1, B_2)$, jeżeli przecięcie zbioru B_{1h} ze zbiorem A^C oraz przecięcie B_{2h} ze zbiorem A są zbiorami pustymi. Jeżeli warunek $B_1 \cap B_2 = \oslash$ nie jest spełniony, to wynik operacji jest zawsze zbiorem pustym.

W praktyce operacje HMT nadaje się szczególnie do wykrywania charakterystycznych punktów na obrazie, takich jak narożniki, krawędzie czy punkty izolowane.

3.2 Wykrywanie krawędzi

Wykrywanie krawędzi stanowi bardzo ważny element procesu analizy obrazu. Informacja o wyróżnionych krawędziach jest często wystarczająca do przeprowadzenia logicznej analizy obrazu. W ogólnym ujęciu wykrywanie krawędzi polega na odnalezieniu pewnych lokalnych nieciągłości w atrybutach obrazu, które odzwierciedlają granice obiektów. Powstają one na skutek wystąpienia właściwości i powierzchni obiektów, cieni itd. Krawędź może być więc zdefiniowana jako skokowa zmiana atrybutów obrazu, zazwyczaj jest to wartość jasności obrazu. Stosuje się m.in. metody polegające na analizie zmian pierwszej i drugiej pochodnej funkcji jasności[28].

Analiza pierwszej pochodnej

Operatory Perwitta i Sobela umożliwiają wyznaczenie dla każdego piksela obrazu estymaty pochodnej w jednym z ośmiu kierunków. Poniżej przedstawiono najczęściej stosowane maski splotu.

Operator Sobela krawędź pozioma:

$$\begin{bmatrix} 1 & 2 & 1 \\ 0 & 0 & 0 \\ -1 & -2 & -1 \end{bmatrix}$$

Operator Sobela krawędź pionowa:

$$\begin{bmatrix} 1 & 0 & -1 \\ 2 & 0 & -2 \\ 1 & 0 & -1 \end{bmatrix}$$

Operator Sobela krawędź ukośna:

$$\begin{bmatrix} 1 & 1 & 1 \\ 0 & 0 & 0 \\ -1 & -1 & -1 \end{bmatrix}$$

Operator Perwitta krawędź pozioma:

$$\begin{bmatrix} 1 & 1 & 1 \\ 0 & 0 & 0 \\ -1 & -1 & -1 \end{bmatrix}$$

Operator Perwitta krawędź pionowa:

$$\begin{bmatrix} 1 & 0 & -1 \\ 1 & 0 & -1 \\ 1 & 0 & -1 \end{bmatrix}$$

Operator Perwitta krawędź ukośna:

$$\begin{bmatrix} 0 & 1 & 1 \\ -1 & 0 & 1 \\ -1 & -1 & 0 \end{bmatrix}$$

Analiza drugiej pochodnej

Do wydobycia krawędzi z obrazu można wykorzystać także operator Laplace'a (laplasjan), aproksymujący drugą pochodną funkcji jasności. W cyfrowej wersji laplasjan dla dowolnego punktu ma postać:

$$L(x, y) = [f(x + 1, y) + f(x - 1, y) + f(x, y - 1) + 4f(x, y)]$$

i można go zaimplementować za pomocą maski. Poniżej pokazano najprostszą implementację o wymiarze 3x3

$$\begin{bmatrix} 0 & 1 & 0 \\ 1 & -4 & 1 \\ 0 & 1 & 0 \end{bmatrix}$$

Niemniej często stosuje się maskę 11x11, w której dodatkowo zrealizowano

splot z rozmyciem gaussowskim [27]

$$\begin{bmatrix} 0 & 0 & 0 & -1 & -1 & 2 & -1 & -1 & 0 & 0 & 0 \\ 0 & 0 & -2 & -4 & -8 & -9 & -8 & -4 & -2 & 0 & 0 \\ 0 & -2 & -7 & -15 & -22 & -23 & -22 & -15 & -7 & -2 & 0 \\ -1 & -4 & -15 & -24 & -14 & -1 & -14 & -24 & -15 & -4 & -1 \\ -1 & -8 & -22 & -14 & 52 & 103 & 52 & -14 & -22 & -8 & -1 \\ -2 & -9 & -23 & -1 & 103 & 178 & 103 & -1 & -23 & -9 & -2 \\ -1 & -8 & -22 & -14 & 52 & 103 & 52 & -14 & -22 & -8 & -1 \\ -1 & -4 & -15 & -24 & -14 & -1 & -14 & -24 & -15 & -4 & -1 \\ 0 & -2 & -7 & -15 & -22 & -23 & -22 & -15 & -7 & -2 & 0 \\ 0 & 0 & -2 & -4 & -8 & -9 & -8 & -4 & -2 & 0 & 0 \\ 0 & 0 & 0 & -1 & -1 & 2 & -1 & -1 & 0 & 0 & 0 \end{bmatrix}$$

Rozdział 4

Implementacja systemu

W rozdziale tym opisano zdobyte doświadczenia oraz napotkane problemy podczas prac nad implementacją własnego systemu segmentacji danych. Z uwagi na badawczy charakter prac uznano, że najefektywniejszą metodologią będzie *projektuj i buduj*. Nie wykonano więc kompletnego projektu aplikacji przed rozpoczęciem prac o charakterze programistycznym, lecz kolejne kroki wyznaczone były przez zdobyte doświadczenia oraz pojawiające się potrzeby. Integralną częścią tego rozdziału jest kod źródłowy, który z powodu znacznej objętości nie mógł zostać w całości zacytowany.

4.1 Wykorzystane komponenty

Program został zaprojektowany jako samodzielna aplikacja (*stand-alone*) z graficznym interfejsem użytkownika. Jako bazę wykorzystano język Java SE7 wraz z wbudowanym pakietem Swing. Prace prowadzono z przy użyciu zintegrowanego środowiska programistycznego *InteliJ IDEA* firmy JetBrains [18]. Istotną konsekwencją tego faktu jest wykorzystanie wbudowanego w to IDE menadżera rozkładu (*layout manager*), który oferuje większe możliwości w stosunku do menadżerów standardowych środowiska Swing. Ponadto użyto *PDFBox*, *Mu PDF*, *OpenCV*.

4.1.1 PDFBox

PDFBox[10] jest darmową biblioteką udostępnioną przez fundację Apache. Umożliwia tworzenie nowych oraz analizowanie i modyfikowanie treści istniejących dokumentów PDF. Posiada rozbudowane API na dwóch poziomach abstrakcji nisko- i wysoko-poziomowe. Przetwarzając dokument można więc operować na typach wprost wynikających ze struktury PDF takich jak liczby, strumienie, słowniki lub na obiektach wirtualnych takich jak strony, zakładki, adnotacje.

Biblioteka w obecnej chwili nie zawiera niestety obsługi wszystkich operatorów wymienionych w oficjalnej specyfikacji [15]. W związku z tym, niektóre strony są prezentowane niepoprawnie - występują na przykład problemy z poprawnym odwzorowaniem barw. Mimo tych braków biblioteka umożliwia formalną analizę dokumentu.

4.1.2 Mu PDF

Ponieważ biblioteka PDFBox nie spełnia wszystkich oczekiwań w zakresie renderowania rastrowego obrazu strony dokumentu do pliku, zdecydowano się na użycie aplikacji alternatywnej.
Mu PDF[5] jest aplikacją dostępną na otwartej licencji Affero GNU GPL służącą do prezentowania dokumentów PDF. Oferuje prostą przeglądarkę z możliwością nawigacji pomiędzy stronami, powiększanie i skalowanie strony, zmiany palety barw (skala szarości oraz inwersja) oraz kontrolą stopnia wygładzania obrazów. Przede wszystkim jednak, zapewnia poprawną obsługę operatorów graficznych, z którymi nie radził sobie PDF Box, a więc możliwe jest wyeksportowanie wszystkich stron dokumentu do pliku rastrowego - między innymi do formatu PNG

4.1.3 OpenCV

OpenCV [22] (Open Source Computer Vision) jest biblioteką funkcji programistycznych do komputerowego przetwarzania obrazów w czasie rzeczywistym. Rozpowszechniana na licencji BSD umożliwia zastosowanie do projek-

tów badawczych oraz komercyjnych. Zaimplementowana jest w języku C++, jednak posiada również interfejsy do C, Pythona i Javy, a wykorzystanie możliwe jest w różnych konfiguracjach sprzętowo-systemowych, włączając w to urządzenia mobilne.

Biblioteka ta posiada szerokie możliwości, a jej funkcje umożliwiają m.in: ogólne przetwarzanie obrazów, wykrywanie oraz rozpoznawanie obiektów, operacje morfologiczne, śledzenie obiektów. Konstrukcja pozwala na wykorzystanie zarówno przez osoby początkujące, które po raz pierwszy stykają się z przetwarzaniem obrazów i chcą poznać tajniki tej sztuki, jak również przez zaawansowanych programistów, dla których biblioteka ta jest tylko jednym z wielu elementów dużego systemu badawczego.

Wybór tej biblioteki był podyktowany następującymi przesłankami:

- dobra struktura danych,
- wszechstronność zastosowań,
- szybkość działania

co potwierdzono w [28].

4.2 Akwizycja danych

Akwizycja danych wejściowych obejmuje dwa etapy:

- pozyskanie pliku PDF,
- wytworzenie z pliku PDF obrazów rastrowych.

4.2.1 Pozyskanie danych

Obecnie istnieje wiele serwisów internetowych umożliwiających dostęp on line do wydań prasowych (np. http://www.e-czasopisma.net.pl/, http://www.e-kiosk.pl/) - odnaleźć je można przy użyciu dowolnej wyszukiwarki internetowej korzystając z frazy *ewydanie* lub *egazeta*.

Większość serwisów tego typu umożliwia - z wykorzystaniem komputera lub urządzenia multimedialnego - na przeglądanie, czytanie, przeszukiwanie treści wg zadanych kryteriów czy sporządzanie notatek. Niemniej serwisy te do prezentacji treści wykorzystują swoje autorskie rozwiązania. W konsekwencji niemożliwe jest pobranie zawartości publikacji w postaci natywnego pliku PDF, gdyż został on już poddany procesowi konwersji do postaci HTML lub JPEG. Nawet jeżeli dane źródłowe są dostarczone do klienta w postaci binarnej, to zazwyczaj są zaszyfrowane (np. przy użyciu mechanizmu Adobe DRM[1]), co również wyklucza ich bezpośrednie użycie.

Niezbędne do przeprowadzania niniejszych badań zbiory danych zostały pozyskane dzięki uprzejmości firmy JaR-eprasa.com, która na co dzień zajmuje się monitoringiem mediów, za co autor składa tej firmie podziękowania.

4.2.2 Konwersja PDF → raster

Na rysunku 4.1 przedstawiono podsystem wytwarzania plików rastrowych.

Rysunek 4.1: Podsystem rasteryzacji:(a) selektor pliku wejściowego, (b) sekcja konfiguracji, (c) panel podglądu, (d) lista wytworzonych stron dokumentu.

Po lewej stronie aplikacji widzimy część konfiguracyjną. Możliwe jest okre-

ślenie takich parametrów jak: ścieżka danych wejściowych (dla uproszczenia przyjęto, że dane wejściowe mogą być zapisane tylko w pliku fizycznym), zakres stron przetwarzania, wybór metody rasteryzacji.
W części środkowej widoczne jest okno podglądu strony. Poniżej niego znajduje się aktualna lista wyrenderowanych zbiorów rastrowych - jeden zbiór odpowiada zazwyczaj jednej stronie dokumentu.

Lista dostępnych metod rasteryzacji jest budowana przy starcie programu. Za pomocą mechanizmu refleksji, odnajdywane są wszystkie klasy implementujące określony interfejs. Taka konstrukcja (stosowana również w kolejnych modułach) daje największe możliwości wzbogacenia aplikacji o nowe implementacje algorytmów czy integrację z alternatywnymi bibliotekami zewnętrznymi bez nadmiernej ingerencji w bazowy kod aplikacji.

PDFBox

Naturalnym wyborem biblioteki do rasteryzacji był PDFBox, co wynikało z tego, że od początku planowano jej wykorzystanie na etapie analizy fizycznej dokumentu. Przeprowadzone testy wykazały jednak niską przydatność tej biblioteki na etapie akwizycji. Główną wadą biblioteki okazały się artefakty widoczne na generowanych obrazach (brak grafiki, tło przysłaniające treści pierwszoplanowe). Ponadto czas przetwarzania dokumentu okazał się wysoce niezadowalający - jest on rzędu kilku sekund na stronę, co w warunkach laboratoryjnych nie budzi zastrzeżeń, natomiast stwarza problemy w użyciu masowym.

Sam proces konwersji pliku PDF do kolekcji obrazów rastrowych przy użyciu PDFBox jest dość prosty. Biblioteka wyposażona jest w dedykowaną temu celowi klasę *PDFImageWritter*. Oprócz możliwości wyboru formatu wyjściowego zbioru, możliwe jest także wybranie konkretnych stron do konwersji, palety barw oraz podanie hasła, jeżeli dokument jest zaszyfrowany przy użyciu standardowych mechanizmów PDF.

MuPDF

Niedogodności opisane powyżej skłoniły do poszukiwania alternatywnej drogi konwersji pdf do obrazu rastrowego. Najpierw z powodzeniem próbowano użyć środowiska Ghostscript[12]. Jest to produkt dojrzały (pierwsza wersja ukazała się w listopadzie 1988 roku [14]) i stabilny. Finalnie wybór jednak padł na sukcesora - program MuPDF[5]. Projekt ten zrodził się z potrzeby odświeżenia interfejsu środowiska Ghostscript, ponadto za priorytetowe aspekty twórcy uznali szybkość działania, możliwie mały rozmiar aplikacji oraz wysoką jakość wytwarzanych obrazów rastrowych.

Program rozpowszechniany jest w postaci aplikacji wykonywalnej typu *exe* w dwóch wersjach: z interfejsem graficznym użytkownika oraz z interfejsem zarządzanym z wiersza poleceń, z którego w niniejszej pracy skorzystano.

Aplikacja posiada szerokie możliwości sterowania przebiegiem generacji obrazu rastrowego. Umożliwia zmianę rozdzielczości, rozmiarów, stopnia wygładzania (antialiasing), generacji plików XML, automatycznego obrotu strony itp.

Porównanie

Wyniki porównania metod rasteryzacji zamieszono w tabeli 4.1. Jako kryteria oceny przyjęto:

- całkowity czas przetwarzania dokumentu,
- średni czas przetwarzania strony,
- bezwzględną liczbę stron przetworzonych błędnie.

Czas mierzono przy pomocy wbudowanych mechanizmów logowania. MuPDF poza czasem całkowitym oraz średnim, prezentował informacje o wynikach skrajnych. W przypadku PDFBox niezbędne były do przeprowadzenia proste operacje arytmetyczne. Jak pokazano, MuPDF osiąga rezultaty lepsze o niemal dwa rzędy wielkości.

Klasyfikacji stron błędnych dokonano poprzez analizę porównawczą - za wzorzec przyjęto obraz, jaki prezentowany jest w wiodącej przeglądarce *Adobe*

dokument	liczba stron	metoda	czas całkowity	czas średni	liczba błędów
Rzeczpospolita 1	20	PDFBox	71.00	3.55	1
		MuPDF	2.47	0.12	0
Wyborcza 1	28	PDFBox	61.00	2.18	1
		MuPDF	3.88	0.14	0
Rzeczpospolita 2	20	PDFBox	83.00	4.15	1
		MuPDF	2.40	0.12	0
Wyborcza 2	28	PDFBox	61.00	2.18	0
		MuPDF	3.12	0.11	0

Tablica 4.1: Porównanie wyników rasteryzacji przy pomocy PDFBox oraz MuPDF. Czasy podane w sekundach.

Reader. Za materiał wzorcowy rozważano przyjęcie skanów z wydruków papierowych, lecz pomysł odrzucono ze względu na to, że proces skanowania z natury rzeczy wprowadza zniekształcenia i szumy, wobec czego określenie na skutek którego ze zjawisk: błędnego działania programu rasteryzacji lub procesu skanowanie wystąpiły błędy, byłoby trudne, a wyniki badań niemiarodajne.

Obrazy rastrowe zapisywane są do plików tymczasowych, stanowiących pośrednie źródło danych do dalszej analizy. Jest to zdeterminowane przez sposób działania MuPDF, który nie umożliwia innego sposobu utrwalenia danych wyjściowych. Może się wydawać, że PDFBox zapewni większą wydajność, ponieważ przetwarzanie wykonywane jest wyłącznie w pamięci operacyjnej, bez konieczności zapisu i następującego chwilę po nim odczytu. Oszacowano jednak, że narzut czasowy na operacje dyskowe jest pomijalnie mały z uwagi na znacznie lepszą wydajność w procesie przetwarzania.

4.3 Analiza fizyczna

4.3.1 Analiza graficzna obrazu rastrowego

Detekcja separatorów rzeczywistych

Na rysunku 4.2 zaprezentowano typowy fragment strony gazety, na którym zaznaczono występujące naturalnie (tzn. widoczne dla czytającego oraz obecne w obrazie rastrowym) separatory. Mają one zazwyczaj formę pionowych lub poziomych linii wyraźnie odróżniających się od tła. Grubość, kolor czy struktura separatora wykazują dużą zmienność, szczególnie jeżeli poddać analizie obrazy pochodzące z różnych plików źródłowych. Podobieństwo jest zazwyczaj zachowane w obrębie danego periodyku.

Opisywane elementy pełnią trojaką rolę: wyznaczają optyczną granicę poszczególnych artykułów widocznych na stronie, oddzielają poszczególne szpalty (kolumny) tekstu w artykule lub są elementem dekoracyjnym, przy czym w niniejszych badaniach istotne są tylko dwa pierwsze.

Rasteryzacja *MuPDF*

W pierwszej fazie badań wykorzystano obrazy rastrowe generowane przez *MuPDF* - który, jak wspomniano w poprzednim podrozdziale, został wybrany jako wiodący system rasteryzacji.

Obraz został wstępnie przetworzony za pomocą filtru rozmywającego *Blur*. Stwierdzono ponadto, że operacja progowania, która mogłaby być użyta na tym etapie, jest niepożądana. W wyniku jej zastosowania, z obrazu w znacznym stopniu usuwane są poszukiwane elementy. W celu wyodrębnienia krawędzi finalnie skorzystano z metody *Canny'ego*, a następnie użyto funkcji *FindContours* w celu wyodrębnienia z obrazu poszczególnych elementów. Dla łatwiejszej interpretacji wyników działania, odnalezione kontury zostały ponownie wykreślone kolorem czerwonym na obrazie wejściowym. Fragment strony pokazano na rysunku 4.3.

Można zauważyć, że kolorem czerwonym oznaczone są niemal wszystkie obiekty widoczne na stronie, wobec czego konieczne okazało się wprowadzenie filtrowania. Klasyfikację kształtów i selekcję zaimplementowano w dwóch kla-

Powrót gladiatora

KAJAKARSTWO | Piotr Siemionowski, główny polski kandydat do złota igrzysk w Londynie, znów może pływać

PAWEŁ WILKOWICZ

Przyzwyczaił już, że robi wszystko inaczej niż koledzy z kadry, więc może to mu po prostu było pisane: w ten weekend w Moskwie wystartował w Pucharze Świata jako jedyny Polak. Bo pozostali kajakarze olimpijczycy są skupieni na przygotowaniach. A dla niego to był dopiero pierwszy start w tym sezonie.

Mistrz świata i Europy w sprincie na 200 m, zdaniem wielu pewny kandydat do medalu w Londynie, leczył przez ostatnie tygodnie mięsień dłoni. W Moskwie wrócił od razu tam, gdzie jest od miesięcy jego miejsce – na podium. Choć jeszcze nie na pierwsze miejsce. – Trzecie jest bardzo dobre, zwłaszcza że czas był dobry. Zwycięzca, Cesar de Cesare z Ekwadoru, powiedział za metą, że ma jednego faworyta w Londynie: Piotrka. Bo po ledwie dwóch tygodniach treningu, a wcześniej miesięcznej przerwie na leczenie przegrał z nim tylko o trzy dziesiąte sekundy. A Cesar mówi, że jest w pełnym gazie – opowiada „Rz" trener Siemionowskiego i kadry sprinterów Mariusz Słowiński.

Wiosło skrojone na miarę

Siemionowski, zbudowany jak gladiator, wiosłujący potężnym, zrobionym na miarę wiosłem, pewny siebie, zapowiadający od dawna, że do Londynu jedzie po złoto, w kwietniu dostał cios od losu. – To było tuż przed próbą krajową, byliśmy na ostatnim treningu przed zawodami, robiliśmy test na 200 m. Piotrek pojechał tak mocno, że zbił mięsień między kciukiem a palcem wskazującym. Długo nie mógł po tym trzymać wiosła. Jeździł do Warszawy do doktora Roberta Śmigielskiego na zastrzyki z komórek macierzystych. Zostawały nam treningi biegowe i siłowe, musieliśmy Piotra wycofywać z kolejnych Pucharów Świata, również tego w Poznaniu. Dopiero od dwóch tygodni trenujemy na sto procent. Najpoważniejszą próbą przed igrzyskami będą mistrzostwa Europy w Zagrzebiu, pod koniec czerwca.

Nic na kredyt

W czasie gdy Siemionowski się leczył, w środowisku kajakarskim zaczęły się dyskusje, czy należy trzymać dla niego miejsce w reprezentacji olimpijskiej bez względu na wszystko, czy też jednak zorganizować krajowe eliminacje, w których mistrz ścigałby się z coraz lepszym Denisem Ambroziakiem.

W kajakarstwie kwalifikacje olimpijskie zdobywa się nie na osobę, ale na osadę, i w sezonie olimpijskim zawsze wybuchają kłótnie, bo każdy mocny klub chciałby wepchnąć do kajaków jak najwięcej swoich zawodników. A Siemionowski długo walczył o niezależność od związku i prawo do indywidualnych treningów, co mu zwolenników wśród działaczy nie przysporzyło.

– Ten konflikt został trochę nakręcony przez działaczy, trenerów innych zawodników. Denis to bardzo zdolny młody chłopak. Ale zawody Pucharu Świata pokazują, że jeszcze nie wytrzymuje presji. To na razie wciąż materiał do obrobienia – mówi Słowiński. – Zresztą nie ma mowy, by Piotr coś dostawał na kredyt. On prowiedział jasno: jak nie będę gotowy na 100 procent, to nie ma sensu jechać do Londynu. Niech wtedy na igrzyska jedzie Denis.

W poprzednich igrzyskach kajakarze zdobyli jeden medal: srebro w kobiecej dwójce Aneta Konieczna-Beata Mikołajczyk.

W Londynie największe szanse będą właśnie w sprincie, który debiutuje w programie igrzysk. Bo równie mocna jak Siemionowski jest Marta Walczykiewicz (wicemistrzyni świata sprzed roku), i ona nie miała problemów ze zdrowiem. Ma je natomiast Aneta Konieczna, która musiała ostatnio przejść zabieg chirurgiczny.

– Anety nigdy nie można skreślać, bo ona ma serce mistrzyni. Jeśli tylko dziewczyny zostawią kłopoty ze zdrowiem za sobą, mogą powalczyć o podium. Świetnie pojechali kwalifikacje kanadyjkarze – mówi Mariusz Słowiński. – Ale realnie patrząc, poważne szanse do medalu są dwie: Piotr i Marta.

♦ **Piotr Siemionowski** był wczoraj trzeci w PŚ w Moskwie

Wygrali bronią rywali

SIATKÓWKA

Nie wygrywali przez 10 lat, a teraz zrobili to drugi raz z rzędu. Polacy pokonali w Lidze Światowej Brazylię 3:2, choć wcześniej stracili sporo nerwów

Dwa tygodnie temu w Toronto reprezentacja Polski przełamała fatalną passę – pokonała mistrzów świata po dziesięciu latach niemocy. Wtedy można było jeszcze narzekać: to nie ta Brazylia co zawsze, bez Murilo, Giby i Leandro Vissotto. Skoro poradziła sobie z nią wtedy Kanada, to dla Polski był to obowiązek.

W niedzielę, w Spodku, było mniej wątpliwości. Do drużyny Bernardo Rezende wrócił kapitan Murilo Endres, Brazylijczycy grali dobrze, ale na Polaków to nie wystarczyło.

To zwycięstwo znaczy więcej, bo Polacy odnieśli je przejętą od rywali bronią: znakomicie bronili, wyprowadzali skuteczne kontry i grali szybko oraz kombinacyjnie. Przez lata Brazylijczycy pokazywali światu, jak grać w ten sposób, a nikt nie potrafił tych zagrywek skopiować. W niedzielę okazało się, że nauczyła się Polska. Trudno o lepsze wiadomości niedługo przed turniejem olimpijskim w Londynie. Polska doczekała się kadry, na której można polegać. Drużyny silnej psychicznie, z czym wcześniej bywało różnie. Takie spotkania jak z Brazylią sprawiają, że spory o to, czy w kadrze powinien grać Mariusz Wlazły czy nie, stają się mało istotne.

Wczoraj w Spodku dobrze grał zwłaszcza Zbigniew Bartman, który nie mylił się, atakując nawet z najtrudniejszych piłek. Rezende mógł tylko wznosić ręce do nieba i skakać przy linii bocznej, a środkowy bloku Brazylijczyków Lucas łapał się za głowę.

Po trzecim secie zaczęły nawet puszczać im nerwy. Brazylijczycy, protestując przeciw decyzji sędziów, wdali się w kłótnię z Polakami, schodząc z boiska. Szczególnie aktywny w tej wymianie zdań był Paweł Zagumny. Sytuację łagodził później Anastasi, klepiąc po plecach libero Brazylijczyków Sergio.

Polacy mogą być zadowoleni z tego zwycięstwa. Na dobre odegnali od siebie kompleks Brazylii, a ponadto pokazali, że są drużyną. Gdy w drugim i trzecim secie zdarzały im się słabsze momenty, Anastasi wpuszczał na boisko rezerwowych: Grzegorza Kosoka i Michała Kubiaka, a oni dodawali swoją cegiełkę.

W beczce miodu jest też łyżka dziegciu. Po raz kolejny zbyt łatwo stracili przewagę i zamiast szybko zakończyć mecz, pozwolili na pięciosetowy maraton. Tie-break to na szczęście był jednak ich popis: wygrali 15:11, a publiczność w Spodku mogła sobie do woli poklaskać.

Brawa należały się za cały weekend w Katowicach. Polacy wygrali wszystkie trzy mecze, w sobotę rozbili Finów 3:0 (26:24, 25:18, 25:18), w piątek również 3:0 Kanadę. A władze PZPS między meczami poinformowały, że kontrakt trenera Anastasiego został przedłużony. Włoch ma prowadzić reprezentację aż do mistrzostw świata w Polsce w 2014 roku. – Mam nadzieję, że podczas tych mistrzostw będę gratulował trenerowi złotego, a przynajmniej srebrnego medalu – mówi prezes PZPS Mirosław Przedpełski.

—Łukasz Majchrzyk

Polska – Brazylia 3:2 (26:24, 23:25, 25:23, 23:25, 15:11).
Finlandia – Kanada 3:2

Tabela:
1. Polska 6 14 17:8 ;
2. Brazylia 6 12 15:11;
3. Kanada 6 6 10:14;
4. Finlandia 6 4 7:16

Rysunek 4.2: Separatory naturalne:(a) (c) naturalne separatory poziome,(b) (d) naturalne separatory pionowe.

Powrót gladiatora

KAJAKARSTWO | Piotr Siemionowski, główny polski kandydat do złota igrzysk w Londynie, znów może pływać

PAWEŁ WILKOWICZ

Przyzwyczaił już, że robi wszystko inaczej niż koledzy z kadry, więc może to mu po prostu było pisane: w ten weekend w Moskwie wystartował w Pucharze Świata jako jedyny Polm. Piotrek pojechał tak mocno, że zbił mięsień między kciukiem a palcem wskazującym. Długo nie mógł po tym trzymać wiosła. Jeździł do Warszawy do doktora Roberta Śmigielskiego na zastrzyki z komórek macierzystych. Zostawały nam treningi biegowe i siłowe, musieliśmy

Rysunek 4.3: Odnalezione kontury na obrazie wejściowym.

syfikatorach, odpowiednio dla separatorów poziomych oraz pionowych. Klasyfikację przeprowadzono badając długość bezwzględną konturu oraz stosunek szerokości do długości. Wartości parametrów dobrano eksperymentalnie. Już pierwsza implementacja dała dość zadowalające rezultaty pokazane na rysunku 4.4, niemniej widoczne są również artefakty w postaci wykreśleń na grafice artykułu.

Wprowadzono więc dodatkowe kryterium odrzucając te, które wykazują zbyt dużą rozpiętość między punktami skrajnymi w kierunku niedominującym.

Na rysunku 4.5 pokazano wynik działania klasyfikatora w wersji ostatecznej. Na tym samym rysunku wskazano także kolorem fioletowym te obiekty, które nie zostały poprawnie sklasyfikowane.
Testy przeprowadzono na próbce 48 stron (Rzeczpospolita1, Wyborcza1), oceniając wizualnie rezultaty. W ujęciu globalnym, skuteczność detekcji separatorów przy użyciu rasteryzacji *MuPdf* wyniosła zaledwie 62%, licząc stosunek separatorów odnalezionych do ogólnej liczby separatorów w całym zbiorze danych. Modyfikacje algorytmu klasyfikatora ani podprocesu wyszukiwania krawędzi (dodatkowe operacje morfologiczne) nie przyniosły oczeki-

Rysunek 4.4: Wynik działania klasyfikatora konturu z widocznymi artefaktami. Kolorem zielonym oznaczono separatory poziome, czerwonym pionowe.

wanej poprawy skuteczności. Z tego powodu należy negatywnie ocenić przydatność tej metody.

Ostry zawał w służbie zdrowia

LECZENIE | Szpitale zwalniają i tną pensje, co spowoduje kolejny kryzys w opiece zdrowotnej

SYLWIA SZPARKOWSKA
EWA ŁOSIŃSKA

W służbie zdrowia tli się konflikt, który może doprowadzić do paraliżu porównywalnego z tym, jaki nastąpił po wejściu w życie ustawy refundacyjnej.

W szpitalach w kraju wybuchają ostre spory. – Nie zajmują opinii publicznej, bo mają lokalny charakter. A napięcie narasta – mówi Iwona Borchulska, szefowa Ogólnopolskiego Związku Zawodowego Pielęgniarek i Położnych. Jutro związek zadecyduje o rozpoczęciu ogólnopolskiej akcji w obronie publicznej służby zdrowia.

Tym razem także chodzi o pieniądze. Dyrektorzy szpitali znaleźli się pod ścianą. Zgodnie z ustawą o lecznictwie muszą do końca roku osiągnąć rentowność, w innym wypadku placówki będą albo dofinansowane przez samorządy, a te nie mają pieniędzy, albo przekształcone w spółki. W tej chwili zobowiązania szpitali, których termin płatności już minął, wynoszą 2,4 mld zł, wszystkie długi placówek przekraczają 10 mld zł. Dyrektorzy szpitali zwalniają pracowników lub wymawiają im dotychczasowe warunki pracy, aby poprawić bilans i uniknąć przekształcenia ich w spółki. Rodzi to konflikty.

W szpitalu MSW w Rzeszowie z pracy rezygnuje połowa lekarzy, wymówienia składają też pielęgniarki. Dyrekcja zgłosiła wniosek o zawieszenie oddziału neurologii, skąd odeszli wszyscy specjaliści. W Skarżysku-Kamiennej w ostatniej chwili udało się uniknąć ewakuacji całej placówki, ale z pracy chcą odejść niemal wszyscy chirurdzy. Do ostrego konfliktu doszło w szpitalu Babińskiego w Krakowie. W jednej z największych placówek psychiatrycznych w kraju trwa głodówka. Rozważana jest ewakuacja 800 pacjentów.

2,4 mld zł
wynoszą obecnie przeterminowane długi szpitali. Zadłużenie wzrasta

Cięcia pensji nakładają się na niezadowolenie pielęgniarek z wydłużenia wieku emerytalnego oraz rozżalenie lekarzy domagających się od kilku miesięcy zmiany prawa dotyczącego przepisywania recept. – Ich protesty w sprawie leków zaktywizowały pozostałe grupy zawodowe – mówi Jerzy Gryglewicz, ekspert ochrony zdrowia z Uczelni Łazarskiego. Dodaje, że protestom sprzyja też narastający konflikt między lekarzami a białym personelem. – Zwiększają się dysproporcje w zarobkach w służbie zdrowia. Szpitale, które płacą za efektywność leczenia, są skłonne znacznie lepiej wynagradzać lekarzy niż pielęgniarki – mówi Gryglewicz. To wszystko sprawia, że służba zdrowia przypomina dziś bombę z opóźnionym zapłonem.

→A7, komentarz →A2,
Rozmowa o opłatach w szpitalach →C6

Zatory biją w firmy z Europy

GOSPODARKA

Najdłużej na uregulowanie należności czekają przedsiębiorcy z Polski, Portugalii i Hiszpanii – wynika z raportu Dun & Bradstreet, międzynarodowej wywiadowni gospodarczej.

Tylko cztery z dziesięciu faktur wystawianych przez europejskie firmy płacone są w terminie. W Polsce w zeszłym roku na czas regulowano przeciętnie co trzeci rachunek. – Szczególnie ostrożnym trzeba być na nowych rynkach i z nowymi partnerami – przyznaje Wojciech Morawski, szef odzieżowej firmy Atlantic.

Najbardziej rzetelnymi kontrahentami, według danych D&B, okazują się firmy niemieckie. To dobra wiadomość dla polskiego biznesu, gdyż jedna czwarta naszego eksportu trafia właśnie za Odrę.

—aft →B2 – 3

EURO 2012 | Gwiazdy już lądują, ale rachunki za stadiony niezapłacone

Do początku mistrzostw już tylko cztery dni. Do Polski zjeżdżają piłkarskie gwiazdy z całej Europy. Jednak nie wiadomo, czy mecze odbędą się bez problemów. Podwykonawcy, którzy pracowali przy budowie Stadionu Narodowego w Warszawie, ostrzegają, że jutro zaczną blokadę obiektu, jeśli NCS nie zagwarantuje im na piśmie, iż zapłaci za wykonane prace jeszcze przed Euro. Szacują, że chodzi o 100 mln zł. Przed stadionem może stanąć 6 tysięcy pracowników wraz ze sprzętem. Dziś między firmami a NCS będą rozmowy ostatniej szansy. We Wrocławiu pikiety podwykonawców, którzy nie dostali pieniędzy, odbyły się w czwartek.

Tymczasem od niedzieli są w Polsce wszyscy nasi grupowi rywale. Czesi przyjechali pociągiem do Wrocławia, Rosjanie i Grecy (na zdjęciu) lądowali w Warszawie – rosyjski czarter z godzinnym opóźnieniem. Dziś przylatują gwiazdy z reprezentacji Niemiec, Portugalii, Holandii i Danii.

W ostatnim meczu przed mistrzostwami Europy reprezentacja Polski pokonała Andorę 4:0.

—ds, dk →A4, →A19 – 20, Opinie →A14

Redaktor prowadzący:
Andrzej Talaga
P R W3 | nakład: 125 159 egz. 35063 X

Rysunek 4.5: Wynik działania klasyfikatora konturu w wersji ostatecznej. Kolorem fioletowym oznaczono separatory, które nie zostały rozpoznane dla rasteryzacji *MuPDF*.

Rasteryzacja *PDFBox*

W fazie drugiej do rasteryzacji wykorzystano *PDFBox*. Mimo niedostatków opisanych w poprzednim rozdziale, biblioteka ta oferuje pełną kontrolę procesu wytwarzania obrazu rastrowego. Podjęto próbę zwiększenia skuteczności wyszukiwania separatorów poprzez poprzez dopasowanie danych wejściowych - obrazu rastrowego.

W procesie wyszukania separatorów naturalnych, elementy tekstowe można potraktować zazwyczaj jako szum informacyjny co wykazano w powyższym podrozdziale. Dlatego w oparciu o silnik biblioteki, zaimplementowano w klasie
PdfBoxRasterizationMethodHideText rasteryzator, który podczas renderowania pomija obiekty tekstowe. Tak znaczne zmniejszenie ilości danych na obrazie rastrowym, wprost przekłada się na wydajność czasową algorytmu, co było pierwszym zaobserwowanym faktem. Przeprowadzając badania na tej samej próbce danych co uprzednio, stwierdzono także znaczne polepszenie skuteczności wykrywania separatorów naturalnych, które w tym wypadku osiągnęło poziom 98.9% - co uznano za w pełni satysfakcjonujące. Na rysunku 4.6 przedstawiono wyniki graficzne.

Rysunek 4.6: Rasteryzacja strony bez elementów tekstowych. Kolorem czerwonym i zielonym wyróżniono odnalezione separatory.

Detekcja separatorów wirtualnych

Po rozpoznaniu separatorów naturalnych przystąpiono do procesu wyszukania separatorów wirtualnych. Jest to pojęcie stworzone na potrzeby niniejszych badań - separator taki definiujemy jako niezadrukowany obszar oddzielający tekst umieszczony w dwóch sąsiednich szpaltach. Na rysunku 4.7 zaprezentowano przykład.

Ostry zawał w służbie zdrowia

LECZENIE | Szpitale zwalniają i tną pensje, co spowoduje kolejny kryzys w opiece zdrowotnej

SYLWIA SZPARKOWSKA
EWA ŁOSIŃSKA

W służbie zdrowia tli się konflikt, który może doprowadzić do paraliżu porównywalnego z tym, jaki nastąpił po wejściu w życie ustawy refundacyjnej.

W szpitalach w kraju wybuchają ostre spory. – Nie zajmują opinii publicznej, bo mają lokalny charakter. A napięcie narasta – mówi Iwona Borchulska, szefowa Ogólnopolskiego Związku Zawodowego Pielęgniarek i Położnych. Jutro związek zadecyduje o rozpoczęciu ogólnopolskiej akcji w obronie publicznej służby zdrowia.

Tym razem także chodzi o pieniądze. Dyrektorzy szpitali znaleźli się pod ścianą. Zgodnie z ustawą o lecznictwie muszą do końca roku osiągnąć rentowność, w innym wypadku placówki będą albo dofinansowane przez samorządy, a te nie mają pieniędzy, albo przekształcone w spółki. W tej chwili zobowiązania szpitali, których termin płatności już minął, wynoszą 2,4 mld zł, wszystkie długi placówek przekraczają 10 mld zł. Dyrektorzy szpitali zwalniają pracowników lub wymawiają im dotychczasowe warunki pracy, aby poprawić bilans i uniknąć przekształcenia ich w spółki. Rodzi to konflikty.

W szpitalu MSW w Rzeszowie z pracy rezygnuje połowa lekarzy, wymówienia składają też pielęgniarki. Dyrekcja zgłosiła wniosek o zawieszenie oddziału neurologii, skąd odeszli wszyscy specjaliści. W Skarżysku-Kamiennej w ostatniej chwili udało się uniknąć ewakuacji całej placówki, ale z pracy chcą odejść niemal wszyscy chirurdzy. Do ostrego konfliktu doszło w szpitalu Babińskiego w Krakowie. W jednej z największych placówek psychiatrycznych w kraju trwa głodówka. Rozważana jest ewakuacja 800 pacjentów.

2,4 mld zł
wynoszą obecnie przeterminowane długi szpitali. Zadłużenie wzrasta

Cięcia pensji nakładają się na niezadowolenie pielęgniarek z wydłużenia wieku emerytalnego oraz rozżalenie lekarzy domagających się od kilku miesięcy zmiany prawa dotyczącego przepisywania recept. – Ich protesty w sprawie leków zaktywizowały pozostałe grupy zawodowe – mówi Jerzy Gryglewicz, ekspert ochrony zdrowia z Uczelni Łazarskiego. Dodaje, że protestom sprzyja też narastający konflikt między lekarzami a białym personelem. – Zwiększają się dysproporcje w zarobkach w służbie zdrowia. Szpitale, które płacą za efektywność leczenia, są skłonne znacznie lepiej wynagradzać lekarzy niż pielęgniarki – mówi Gryglewicz. To wszystko sprawia, że służba zdrowia przypomina dziś bombę z opóźnionym zapłonem.

→A7, komentarz →A2,
Rozmowa o opłatach w szpitalach →C6

Rysunek 4.7: Artykuł z oznaczonymi na niebiesko separatorami wirtualnymi.

Problem detekcji separatorów wirtualnych określono jako zadanie wyszukiwania białych linii w obrazie binarnym. Obraz wejściowy po przetwarzaniu wstępnym zawierać powinien czarne prostokąty w tych fragmentach, w których zawarta jest treść drukowana. W celu ułatwienia procesu binaryzacji oraz na bazie doświadczeń zdobytych przy detekcji separatorów naturalnych, od razu przystąpiono do implementacji dedykowanego rasteryzatora, z pominięciem analizy obrazu naturalnego. Został on zapisany w klasie *PdfBoxRasterizationMethodBlackRectanlgeText*. W procesie renderowania strony, zamiast glifu litery wykreślany jest wypełniony prostokąt otaczający w kolorze czarnym. Dzięki takiemu zabiegowi obraz wejściowy jest bardziej spójny.

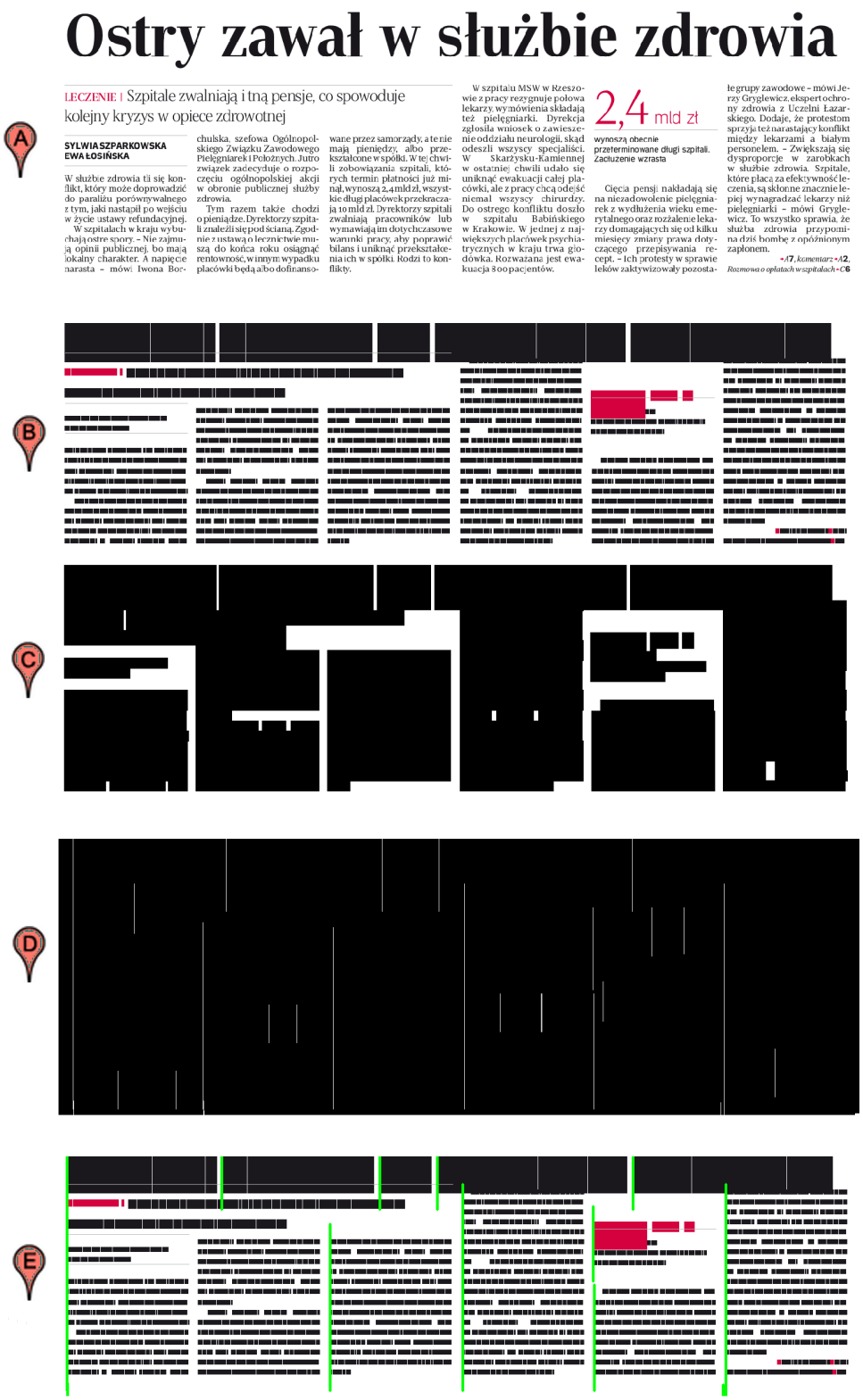
Proces przetwarzania wstępnego składa się z kilku operacji: progowania binarnego, erozji morfologicznej oraz operacji otwarcia. Operacje morfologiczne wykonywane są z wykorzystaniem dedykowanego elementu struktural-

nego, którego parametry zostały określone doświadczalnie. Kluczowym parametrem jest średnia wysokość linijki tekstu, która w prezentowanym przykładzie wynosiła 9px. W operacjach erozji oraz otwarcia, odpowiednio zastosowano elementy strukturalne o rozmiarach (4x12) oraz (3x17)px. Operacja erozji powoduje *zlanie* widocznych linijek tekstu z ich poziomymi separatorami.

Operacja otwarcia ma mniej kluczowe znaczenia dla skuteczności samego procesu, natomiast pozytywnie wpływa na wydajność całego algorytmu eliminując z prostokątnych obszarów tła lokalne białe plamy.

Przetwarzanie właściwe oparte jest o operator *Hit-Or-Miss*. Biblioteka *OpenCV* nie zawiera implementacji tego operatora, wobec czego została ona dokonana w klasie *HitOrMissUtil* na podstawie informacji zaczerpniętych z internetu [20]. Poszukiwany element strukturalny można opisać jako pionową krawędź. Dla uwypuklenia odnalezionych linii oraz złączenia w całość linii, które zostały wykryte fragmentarycznie, obraz został poddany operacji dylacji z elementem strukturalnym (1x25)px. W kroku następnym linie pionowe zostały wykryte przy pomocy wbudowanej funkcji odnajdywania konturów.

Podobnie jak miało to miejsce w procesie wyszukiwania separatorów naturalnych, tak i tutaj koniczne było wdrożenie dodatkowej selekcji odnalezionych elementów. Przyjęto analogiczne kryterium kształtu oraz rozmiaru jak wcześniej.
Dane z poszczególnych etapów procesu zgromadzono i zaprezentowano na rysunku 4.8.



Ostry zawał w służbie zdrowia

LECZENIE | Szpitale zwalniają i tną pensje, co spowoduje kolejny kryzys w opiece zdrowotnej

SYLWIA SZPARKOWSKA
EWA ŁOSIŃSKA

W służbie zdrowia tli się konflikt, który może doprowadzić do paraliżu porównywalnego z tym, jaki nastąpił po wejściu w życie ustawy refundacyjnej.

W szpitalach w kraju wybuchają ostre spory. – Nie zajmują opinii publicznej, bo mają lokalny charakter. A napięcie narasta – mówi Iwona Borchulska, szefowa Ogólnopolskiego Związku Zawodowego Pielęgniarek i Położnych. Jutro związek zadecyduje o rozpoczęciu ogólnopolskiej akcji w obronie publicznej służby zdrowia.

Tym razem także chodzi o pieniądze. Dyrektorzy szpitali znaleźli się pod ścianą. Zgodnie z ustawą o lecznictwie muszą do końca roku osiągnąć rentowność, w innym wypadku placówki będą albo dofinansowane przez samorządy, a te nie mają pieniędzy, albo przekształcone w spółki. W tej chwili zobowiązania szpitali, których termin płatności już minął, wynoszą 2,4 mld zł, wszystkie długi placówek przekraczają 10 mld zł. Dyrektorzy szpitali zwalniają pracowników lub wymawiają im dotychczasowe warunki pracy, aby poprawić bilans i uniknąć przekształcenia ich w spółki. Rodzi to konflikty.

W szpitalu MSW w Rzeszowie z pracy rezygnuje połowa lekarzy, wymówienia składają też pielęgniarki. Dyrekcja zgłosiła wniosek o zawieszenie oddziału neurologii, skąd odeszli wszyscy specjaliści. W Skarżysku-Kamiennej w ostatniej chwili udało się uniknąć ewakuacji całej placówki, ale z pracy chcą odejść niemal wszyscy chirurdzy. Do ostrego konfliktu doszło w szpitalu Babińskiego w Krakowie. W jednej z największych placówek psychiatrycznych w kraju trwa głodówka. Rozważana jest ewakuacja 800 pacjentów.

2,4 mld zł
wynoszą obecnie przeterminowane długi szpitali. Zadłużenie wzrasta

Cięcia pensji nakładają się na niezadowolenie pielęgniarek z wydłużenia wieku emerytalnego oraz rozżalenie lekarzy domagających się od kilku miesięcy zmiany prawa dotyczącego przepisywania recept. – Ich protesty w sprawie leków zaktywizowały pozostałe grupy zawodowe – mówi Jerzy Gryglewicz, ekspert ochrony zdrowia z Uczelni Łazarskiego. Dodaje, że protestom sprzyja też narastający konflikt między lekarzami a białym personelem. – Zwiększają się dysproporcje w zarobkach w służbie zdrowia. Szpitale, które płacą za efektywność leczenia, są skłonne znacznie lepiej wynagradzać lekarzy niż pielęgniarki – mówi Gryglewicz. To wszystko sprawia, że służba zdrowia przypomina dziś bombę z opóźnionym zapłonem.

→A7, komentarz →A2,
Rozmowa o opłatach w szpitalach →C6

Rysunek 4.8: Poszczególne fazy detekcji separatorów wirtualnych: a - oryginał, b - rasteryzacja dedykowana, c - przetwarzanie wstępne, d - przetwarzanie właściwe, e - wynik końcowy

4.3.2 Analiza zawartości natywnej PDF

W kolejnym kroku przystąpiono do analizy natywnej zawartości pliku PDF. Pozyskanie informacji w przypadku pliku PDF jest możliwe tylko w jeden sposób - parsowanie zawartości. Biblioteka *PDFBox* dostarcza do tego celu wygodne, wysokopoziomowe API. Korzystając z klasy *PDFStreamEngine* programista odpowiedzialny jest jedynie za dostarczenie procedur obsługi poszczególnych operatorów, natomiast zaniedbać może proces analizy leksykalnej i składniowej. Biblioteka zadba o odpowiednią klasyfikację danych wejściowych, zgromadzenie aktualnych parametrów i przekazanie jej do procedury obsługi operatora.

Procedury obsługi operatora zorganizowane są w postaci hierarchii klas dziedziczących po abstrakcyjnej klasie *OperatorProcessor*. Ponieważ każdy z operatorów może posiadać dowolną liczbę klas implementujących jego obsługę, na etapie konstrukcji obiektu *PDFStreamEngine* definiowane jest oczekiwane mapowanie. Niektóre z operatorów mogą być również oznaczone jako nieobsługiwane. Zabieg ten ma dwojakie zastosowanie - poprawia ogólną wydajność przetwarzania (np. poprzez pominięcie operacji nie mających wpływu na dane wyjściowe jak renderowanie grafiki w procesie ekstrakcji tekstu), oraz umożliwia (częściowe) przetworzenie plików zawierających niezaimplementowane jeszcze w bibliotece operatory.

Ekstrakcja elementów tekstowych

Operatory, które zaangażowane są w obsługę obiektów tekstowych, podzielić można na kilka grup. Pierwszą z nich stanowi zbiór dwuelementowy *BT, ET* - są to operatory najwyższego poziomu kolejno definiujące początek i koniec obiektu tekstowego.

Grupa druga *Td, TD, Tm, T** zawiera operatory odpowiedzialne za pozycjonowanie obiektu tekstowego na urządzeniu wyjściowym. Pozycjonowanie może mieć charakter względny bądź absolutny.

Trzecia grupa *Tc, Tw, Tz, TL, Tm, Tf, Tr, Ts* obejmuje operacje związane z określeniem parametrów renderowanego tekstu takich jak wielkość czcionki, odstęp między literami, odstęp między wyrazami, skalowanie itp.

Ostatnią grupę stanowią operatory renderowania *Tj, TJ, ′, ″*, których zadaniem jest fizyczne wykreślenie glifów na urządzeniu wyjściowym.

Pozyskanie użytecznych informacji o obiektach tekstowych wymaga poprawnej i aktywnej implementacji wymienionych operatorów w procesie parsowania danych wejściowych. Wymagania te zostały zrealizowane w klasie *TextPositionExctractor*, która jest rozszerzeniem standardowego parsera biblioteki - klasy *PDFStreamEngine*. Na podstawie wyników badań określono, że implementacja operatorów tekstowych dostarczona z biblioteką jest wystarczająco dobra. Ponadto określono wpływ deaktywacji operatorów innych niż wyżej wymienione na szybkość działania modułu - średni czas przetwarzania pliku zmniejszył się sześćdziesięciokrotnie, co należy uznać za znaczące. Zebrane dane przedstawiono w tabeli 4.2.

dokument	tylko tekstowe	wszystkie
Rzeczpospolita 1	1,3	71,0
Wyborcza 1	0,9	61,0
Rzeczpospolita 2	1,2	83,0
Wyborcza 2	1,0	61,0

Tablica 4.2: Porównanie czasu ekstrakcji obiektów tekstowych przy deaktywacji operatorów graficznych. Czasy podane w sekundach.

Ekstrakcja separatorów treści

Działania podjęte w celu wyodrębnienia bezpośrednio z pliku PDF separatorów, które wskazano na rysunku 4.2, nie przyniosły oczekiwanych efektów. Główna trudność polegała na tym, że w przetwarzanych plikach separatory nawet w obrębie jednego artykułu były renderowane przy pomocy różnych technik: jako ogólne obiekty graficzne (np. krzywe Béziera), jako specyficzne obiekty tekstowe lub wykorzystaniem zewnętrznych obrazów rastrowych ewentualnie wektorowych. Ponieważ przy wykorzystaniu tych samych technik renderowane są również inne obiekty, niemożliwe było stworzenie klasyfikatora opartego na rodzaju wykonywanej operacji.

Z uwagi na wskazane problemy oraz pozytywne wyniki analizy rastrowej, zaprzestano dalszych badań w opisywanym kierunku.

4.3.3 Proces grupowania elementów tekstowych

Dysponując informacjami o obiektach tekstowych zawartych w dokumencie, przystąpiono do procesu wyodrębniania obszarów homogenicznych. W niniejszym podrozdziale opisano kolejne wersje algorytmu wraz z procesem ewolucji.

Za punkt wyjścia przyjęto, że z pliku źródłowego PDF została wyeksportowana lista obiektów tekstowych. W celu uniezależnienia się od API biblioteki PDFBox oraz zdefiniowanych w niej typów danych, stworzono klasę *TextObjectWrapper*, której zadaniem jest przechowywanie danych o obiekcie tekstowym takich jak lokalizacja, obszar otaczający, dane czcionki. Użycie w projekcie innej biblioteki niż PDFBox będzie wymagało jedynie dopisania stosownego konwertera, który na podstawie natywnego typu danych będzie potrafił kreować obiekty opisywanej klasy.

Wersja 1 - tylko metadane tekstowe

Pierwsza, najprostsza wersja algorytmu grupowania elementów tekstowych bazuje na metodach opisanych w artykule [9]. W badaniach autorzy wykorzystali alternatywną bibliotekę do analizy plików PDF firmy Datalogic. Z uwagi na jej komercyjny charakter, niemożliwe było bezpośrednie porównanie działania z biblioteką PDFBox.

Istotna różnica między tymi bibliotekami tkwi w tym, że PDBox tworzy obiekt opakowujący dla każdego, pojedynczego znaku z obiektu tekstowego, podczas gdy biblioteka alternatywna tworzy je dla całych słów. Innymi słowy PDFBox zwraca dane na większym poziomie szczegółowości.

Autorzy artykułu zaproponowali użycie następującego operatora opisującego relację między dwiema nieujemnymi wartościami.

$$\Delta(v_1, v_2) \begin{cases} 0, & gdy\ v_1 = 0\ oraz\ v_2 = 0 \\ \infty & gdy\ (v_1 \cdot v_2)\ oraz\ v_1 \neq v_2 \\ |v_1 - v_2|/min(v_1, v_2) & w \text{ pozostałych wypadkach} \end{cases}$$

Operator ten został użyty do zdefiniowania następujących reguł formowania linii tekstu:

- Odstęp pionowy; odstęp w kierunku pionowym pomiędzy dwoma obszarami otaczającymi powinien być dostatecznie duży i spełniać zależność

  $$O(q_i, q_j) > k_o \cdot min(h_i, h_j),$$

  gdzie O jest odstępem w kierunku pionowym, h wysokością obiektu, k_o parametrycznym współczynnik progowy,

- Rozmiar czcionki; różnica rozmiaru czcionki między dwoma obiektami powinna być dostatecznie mała i spełniać zależność

  $$\Delta(f_i, f_j) < k_{fh},$$

  gdzie f jest rozmiarem czcionki, k_{fh} parametrycznym współczynnikiem progowym,

- Odstęp poziomy; odstęp poziomy powinien być dostatecznie mały i spełniać zależność

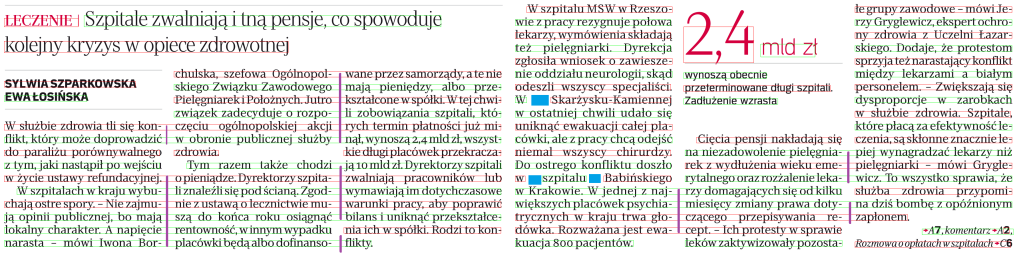
  $$d_{ij} < k_{dq} \cdot min(f_i, f_j),$$

  gdzie f jest rozmiarem czcionki, d_{ij} odstępem między obszarami otaczającymi, k_{dq} parametrycznym współczynnikiem progowym.

Autorzy użyli średniej ważonej dla wyliczenia wartości rozmiaru czcionki oraz współrzędnej pionowej. Ponieważ PDFBox tworzy obiekty dla pojedynczych znaków, operacja ta została pominięta. Algorytm grupowana działa więc na większym poziomie ogólności.

Podobne zależności wykorzystywane są dla łączenia linii tekstu w większe bloki. W tym wypadku istotna jest wzajemna odległość pomiędzy liniami tekstu mierzona w kierunku pionowym oraz wysokość czcionki. W każdej iteracji analizie podlegają aż trzy linie tekstu i ich wzajemne relacje, dzięki czemu możliwe jest lepsze określenie przynależności linii leżących na granicy dwóch obszarów.

Niestety określenie optymalnych wartości wyżej opisywanych współczynników często napotyka na trudności lub jest niemożliwe. Jak pokazano na rysunku 4.9, nadmierne zwiększanie parametru odstępu poziomego prowadzi do zwiększenia liczby błędów polegających na złączeniu w logiczną linię tekstu fragmentów przynależnych do odrębnych szpalt (oznaczone kolorem fioletowym). W prezentowanym przykładzie, wartość tego parametru jest na tyle duża, aby powodować błędy złączeń między szpaltami a jednocześnie zbyt mała, aby zapewnić poprawne złączenie elementów w linię.


LECZENIE | Szpitale zwalniają i tną pensje, co spowoduje kolejny kryzys w opiece zdrowotnej

SYLWIA SZPARKOWSKA
EWA ŁOSIŃSKA

W służbie zdrowia tli się konflikt, który może doprowadzić do paraliżu porównywalnego z tym, jaki nastąpił po wejściu w życie ustawy refundacyjnej.

W szpitalach w kraju wybuchają ostre spory. – Nie zajmują opinii publicznej, bo mają lokalny charakter. A napięcie narasta – mówi Iwona Borchulska, szefowa Ogólnopolskiego Związku Zawodowego Pielęgniarek i Położnych. Jutro związek zadecyduje o rozpoczęciu ogólnopolskiej akcji w obronie publicznej służby zdrowia.

Tym razem także chodzi o pieniądze. Dyrektorzy szpitali znaleźli się pod ścianą. Zgodnie z ustawą o lecznictwie muszą do końca roku osiągnąć rentowność, w innym wypadku placówki będą albo dofinansowane przez samorządy, a te nie mają pieniędzy, albo przekształcone w spółki. W tej chwili zobowiązania szpitali, których termin płatności już minął, wynoszą 2,4 mld zł, wszystkie długi placówek przekraczają 10 mld zł. Dyrektorzy szpitali zwalniają pracowników lub wymawiają im dotychczasowe warunki pracy, aby poprawić bilans i uniknąć przekształcenia ich w spółki. Rodzi to konflikty.

W szpitalu MSW w Rzeszowie z pracy rezygnuje połowa lekarzy, wymówienia składają też pielęgniarki. Dyrekcja zgłosiła wniosek o zawieszenie oddziału neurologii, skąd odeszli wszyscy specjaliści. W Skarżysku-Kamiennej w ostatniej chwili udało się uniknąć ewakuacji całej placówki, ale z pracy chcą odejść niemal wszyscy chirurdzy. Do ostrego konfliktu doszło w szpitalu Babińskiego w Krakowie. W jednej z największych placówek psychiatrycznych w kraju trwa głodówka. Rozważana jest ewakuacja 800 pacjentów.

2,4 mld zł
wynoszą obecnie przeterminowane długi szpitali. Zadłużenie wzrasta

Cięcia pensji nakładają się na niezadowolenie pielęgniarek z wydłużenia wieku emerytalnego oraz rozżalenie lekarzy domagających się od kilku miesięcy zmiany prawa dotyczącego przepisywania recept. – Ich protesty w sprawie leków zaktywizowały pozostałe grupy zawodowe – mówi Jerzy Gryglewicz, ekspert ochrony zdrowia z Uczelni Łazarskiego. Dodaje, że protestom sprzyja też narastający konflikt między lekarzami a białym personelem. – Zwiększają się dysproporcje w zarobkach w służbie zdrowia. Szpitale, które płacą za efektywność leczenia, są skłonne znacznie lepiej wynagradzać lekarzy niż pielęgniarki – mówi Gryglewicz. To wszystko sprawia, że służba zdrowia przypomina dziś bombę z opóźnionym zapłonem.

•A7, komentarz •A2,
Rozmowa o opłatach w szpitalach •C6

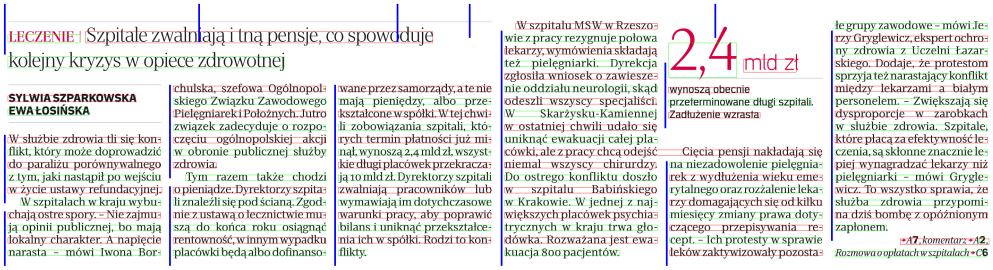
Rysunek 4.9: Przykład działania algorytmu prostego. Kolorem niebieskim oznaczono błędy w rozpoznaniu linii, kolorem fioletowym błędne sklejenia międzyszpaltowe.

Wersja 2 - z wykorzystaniem separatorów

W toku niniejszych badań, zaproponowano rozszerzenie wyżej opisanego algorytmu o dodatkowe reguły obostrzające warunki grupowania elementów w oparciu o naturalne i wirtualne separatory. O ile pomiędzy obiektami, które kandydują do połączenia w obiekt logiczny, umieszczony jest separator

dowolnego typu, to operacja złączenia jest zabroniona. Reguła ta posiada najwyższy priorytet w drzewie decyzyjnym.

Na rysunku 4.10 pokazano wynik grupowania dla tego samego artykułu co powyżej po uwzględnieniu opisywanego rozszerzenia algorytmu. Jednocześnie zwiększając wartość parametru odstępu poziomego udało się poprawnie rozpoznać wszystkie linie tekstu.


LECZENIE | Szpitale zwalniają i tną pensje, co spowoduje kolejny kryzys w opiece zdrowotnej

SYLWIA SZPARKOWSKA
EWA ŁOSIŃSKA

W służbie zdrowia tli się konflikt, który może doprowadzić do paraliżu porównywalnego z tym, jaki nastąpił po wejściu w życie ustawy refundacyjnej.

W szpitalach w kraju wybuchają ostre spory. – Nie zajmują opinii publicznej, bo mają lokalny charakter. A napięcie narasta – mówi Iwona Borchulska, szefowa Ogólnopolskiego Związku Zawodowego Pielęgniarek i Położnych. Jutro związek zadecyduje o rozpoczęciu ogólnopolskiej akcji w obronie publicznej służby zdrowia.

Tym razem także chodzi o pieniądze. Dyrektorzy szpitali znaleźli się pod ścianą. Zgodnie z ustawą o lecznictwie muszą do końca roku osiągnąć rentowność, w innym wypadku placówki będą albo dofinansowane przez samorządy, a te nie mają pieniędzy, albo przekształcone w spółki. W tej chwili zobowiązania szpitali, których termin płatności już minął, wynoszą 2,4 mld zł, wszystkie długi placówek przekraczają 10 mld zł. Dyrektorzy szpitali zwalniają pracowników lub wymawiają im dotychczasowe warunki pracy, aby poprawić bilans i uniknąć przekształcenia ich w spółki. Rodzi to konflikty.

W szpitalu MSW w Rzeszowie z pracy rezygnuje połowa lekarzy, wymówienia składają też pielęgniarki. Dyrekcja zgłosiła wniosek o zawieszenie oddziału neurologii, skąd odeszli wszyscy specjaliści. W Skarżysku-Kamiennej w ostatniej chwili udało się uniknąć ewakuacji całej placówki, ale z pracy chcą odejść niemal wszyscy chirurdzy. Do ostrego konfliktu doszło w szpitalu Babińskiego w Krakowie. W jednej z największych placówek psychiatrycznych w kraju trwa głodówka. Rozważana jest ewakuacja 800 pacjentów.

2,4 mld zł

wynoszą obecnie przeterminowane długi szpitali. Zadłużenie wzrasta

Cięcia pensji nakładają się na niezadowolenie pielęgniarek z wydłużenia wieku emerytalnego oraz rozżalenie lekarzy domagających się od kilku miesięcy zmiany prawa dotyczącego przepisywania recept. – Ich protesty w sprawie leków zaktywizowały pozosta-

le grupy zawodowe – mówi Jerzy Gryglewicz, ekspert ochrony zdrowia z Uczelni Łazarskiego. Dodaje, że protestom sprzyja też narastający konflikt między lekarzami a białym personelem. – Zwiększają się dysproporcje w zarobkach w służbie zdrowia. Szpitale, które płacą za efektywność leczenia, są skłonne znacznie lepiej wynagradzać lekarzy niż pielęgniarki – mówi Gryglewicz. To wszystko sprawia, że służba zdrowia przypomina dziś bombę z opóźnionym zapłonem.

→A7, komentarz →A2, Rozmowa o opłatach w szpitalach →C6

Rysunek 4.10: Przykład działania algorytmu zmodyfikowanego.

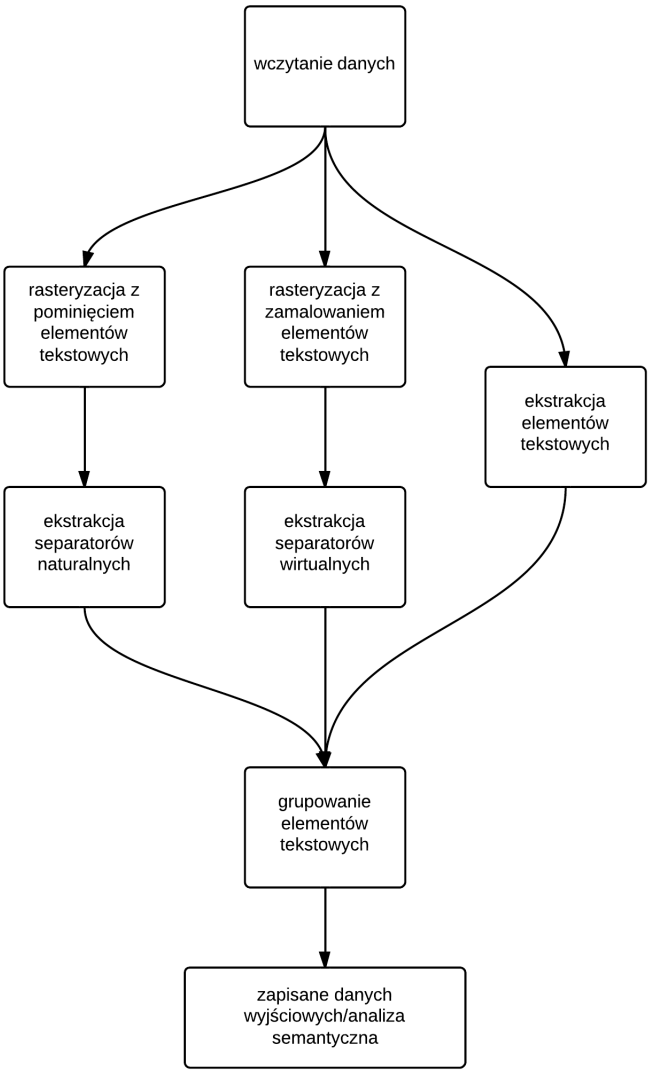
Podsumowanie

Na rysunku 4.11 przedstawiono końcowy, wypracowany algorytm analizy fizycznej. Jak już opisano powyżej, w celu uzyskania jak największej skuteczności działania, operacja rasteryzacji wykonywana jest dwukrotnie - w sposób optymalny dla rodzaju wyszukiwanych separatorów. Informacje o separatorach są następnie poddawane dalszej analizie łącznie z danymi o natywnych elementach tekstowych.

4.4 Wyniki testów

W celu wykazania poprawności działania zaproponowanego rozwiązania przeprowadzono testy porównawcze, których wyniki zaprezentowano w tabeli 4.3.

Jako główne kryterium przyjęto dwa wskaźniki: liczbę błędnie rozpoznanych linii (linia rozpoznana jest fragmentarycznie) oraz liczba sklejeń po-



Rysunek 4.11: Wypracowany algorytm przetwarzania dokumentu

między szpaltami. Dla wariantu prostego algorytmu, wartości parametrów dobrano tak, aby zminimalizować liczbę błędów w liniach i jednocześnie nie generować żadnych błędów w szpaltach. Z punktu widzenia błędów szpaltowych są to więc warunki graniczne.
Dla wariantu modyfikowanego algorytmu zwiększano wartości parametrów sterujących obserwując sukcesywne zmniejszanie się liczby błędów w liniach. Jednocześnie zaobserwowano nieznaczną liczbę błędów w szpaltach, czego przyczyny opisano poniżej.

Na rysunku 4.12 zaprezentowano przykład rozpoznawania linii tekstu artykułu z wykorzystaniem algorytmu prostego (część a) oraz modyfikowanego (część b). Jak pokazano, cztery linie tekstu nie zostały poprawnie rozpoznane jako linie pełne - w ich obrębie znajdują się odstępy. Powyższy błąd nie wystąpił w przypadku, gdy użyty został algorytm modyfikowany -po wcześniejszym przeprowadzeniu operacji detekcji separatorów (w tym wypadku wystarczające było określenie separatorów wirtualnych).

W celu zminimalizowania ewentualnych błędów pomiarowych, ocena działania algorytmu i detekcja błędów wykonywana była manualnie na podstawie wyników rastrowych. Jest to zajęcie niezwykle żmudne i czasochłonne, czym należy tłumaczyć relatywnie niską ilość przebadanego materiału, która jednak była wystarczająca dla udowodnienia poprawności postawionej hipotezy.

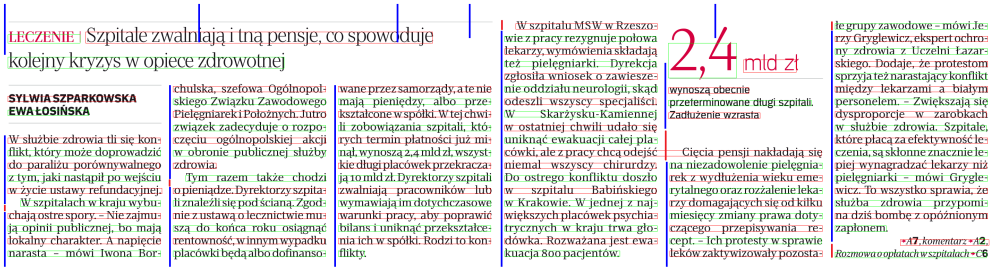
dokument	stron	algorytm	liczba błędów w liniach	liczba błędów w szpaltach
Rzeczpospolita	6	prosty	28	-
		modyfikowany	0	4
Wyborcza	4	prosty	21	-
		modyfikowany	0	6

Tablica 4.3: Porównanie wyników działania algorytmu rozpoznawania linii.

Rysunek 4.12: Graficzne wyniki działania: (a) algorytm prosty - niebieskim znacznikiem oznaczono linie rozpoznane błędnie; (b) algorytm modyfikowany -niebieskim kolorem wykreślono separatory wirtualne

Sklejenia międzyszpaltowe

Pojawienie się sklejeń międzyszpaltowych wytłumaczyć można niedoskonałością procesu wyszukiwania separatorów wirtualnych. Przykład zaprezentowano na rysunku 4.13.


LECZENIE | Szpitale zwalniają i tną pensje, co spowoduje kolejny kryzys w opiece zdrowotnej

SYLWIA SZPARKOWSKA
EWA ŁOSIŃSKA

W służbie zdrowia tli się konflikt, który może doprowadzić do paraliżu porównywalnego z tym, jaki nastąpił po wejściu w życie ustawy refundacyjnej.

W szpitalach w kraju wybuchają ostre spory. – Nie zajmują opinii publicznej, bo mają lokalny charakter. A napięcie narasta – mówi Iwona Borchulska, szefowa Ogólnopolskiego Związku Zawodowego Pielęgniarek i Położnych. Jutro związek zadecyduje o rozpoczęciu ogólnopolskiej akcji w obronie publicznej służby zdrowia.

Tym razem także chodzi o pieniądze. Dyrektorzy szpitali znaleźli się pod ścianą. Zgodnie z ustawą o lecznictwie muszą do końca roku osiągnąć rentowność, w innym wypadku placówki będą albo dofinansowane przez samorządy, a te nie mają pieniędzy, albo przekształcone w spółki. W tej chwili zobowiązania szpitali, których termin płatności już minął, wynoszą 2,4 mld zł, wszystkie długi placówek przekraczają 10 mld zł. Dyrektorzy szpitali zwalniają pracowników lub wymawiają im dotychczasowe warunki pracy, aby poprawić bilans i uniknąć przekształcenia ich w spółki. Rodzi to konflikty.

W szpitalu MSW w Rzeszowie z pracy rezygnuje połowa lekarzy, wymówienia składają też pielęgniarki. Dyrekcja zgłosiła wniosek o zawieszenie oddziału neurologii, skąd odeszli wszyscy specjaliści. W Skarżysku-Kamiennej w ostatniej chwili udało się uniknąć ewakuacji całej placówki, ale z pracy chcą odejść niemal wszyscy chirurdzy. Do ostrego konfliktu doszło w szpitalu Babińskiego w Krakowie. W jednej z największych placówek psychiatrycznych w kraju trwa głodówka. Rozważana jest ewakuacja 800 pacjentów.

2,4 mld zł
wynoszą obecnie przeterminowane długi szpitali. Zadłużenie wzrasta

Cięcia pensji nakładają się na niezadowolenie pielęgniarek z wydłużenia wieku emerytalnego oraz rozżalenie lekarzy domagających się od kilku miesięcy zmiany prawa dotyczącego przepisywania recept. – Ich protesty w sprawie leków zaktywizowały pozosta-

le grupy zawodowe – mówi Jerzy Gryglewicz, ekspert ochrony zdrowia z Uczelni Łazarskiego. Dodaje, że protestom sprzyja też narastający konflikt między lekarzami a białym personelem. – Zwiększają się dysproporcje w zarobkach w służbie zdrowia. Szpitale, które płacą za efektywność leczenia, są skłonne znacznie lepiej wynagradzać lekarzy niż pielęgniarki – mówi Gryglewicz. To wszystko sprawia, że służba zdrowia przypomina dziś bombę z opóźnionym zapłonem.

→A7, komentarz →A2.
Rozmowa o opłatach w szpitalach →C6

Rysunek 4.13: Wyszukiwanie separatorów wirtualnych. Kolorem czerwonym oznaczono miejsca nieciągłości.

W takich sytuacjach jak brzeg strony, czy zwiększone wcięcie akapitowe może się zdarzyć, że separator wirtualny będzie nieciągły, wskutek czego nie zapewni ochrony przed sklejeniem międzyszpaltowym dla wszystkich wierszy. W większości wypadków samo wcięcie akapitowe nie jest przyczyną powstania błędu sklejenia, natomiast jego zbieg ze zwiększonym odstępem pionowym czy sąsiadująca infografika, może do tego doprowadzić, co również widoczne jest na zaprezentowanym przykładzie.

Udoskonalenie algorytmu wykrywania separatorów wirtualnych poprzez ich zobojętnienie na obecność wcięć akapitowych powinno poprawić skuteczność działania.

Rozdział 5

Podsumowanie i wnioski

W części teoretycznej pracy scharakteryzowano języki opisu strony oraz opisano podstawowe cechy formatu PDF. Przedstawiono najważniejsze jego aspekty, które mają wpływ na temat przewodni pracy - ekstrakcję treści i detekcję artykułów. Przedstawiono proces tworzenia dokumentów i wskazano jego najistotniejsze fazy. Omawiane w pracy zagadnienia segmentacji danych z pliku PDF są dosyć mało popularne w literaturze, niemniej opisano te źródła, do których zdołano dotrzeć. Ponadto przedstawiono teoretyczne podstawy wybranych metod przetwarzania obrazów - jedynie tych, które zostały wykorzystane w pracy.

W części praktycznej przygotowano kompletny program do prowadzenia analizy działania algorytmów rasteryzacji oraz analizy fizycznej plików PDF. Zaimplementowano rasteryzatory z wykorzystaniem dwóch zewnętrznych bibliotek, które zastępują popularne czytniki PDF. Ponadto opracowano i uruchomiono dwa rasteryzatory specjalizowane, których działanie ma kluczowy wpływ na proces analizy graficznej.

Proces analizy zawartości prowadzono dwutorowo: poprzez analizę danych rastrowych z wykorzystaniem bibliotek przetwarzania obrazu oraz poprzez analizę danych tekstowych dostępnych natywnie z pliku PDF. Postawiono hipotezę, że połączenie tych dwóch metod i wykorzystanie ich wyników w sposób łączny pozwala na zwiększenie skuteczności działania w procesie analizy i ekstrakcji danych tekstowych z pliku PDF.

W pracy zaimplementowano jeden z opisanych w literaturze algorytmów analizy tekstu plików PDF. Następnie algorytm został tak zmodyfikowany, aby uwzględnić dane pozyskane z analizy graficznej. Przeprowadzone na końcu badania udowodniły słuszność postawionych postulatów.

Zasadniczym punktem badań, który przesądził zresztą o ich pozytywnym zakończeniu, było rozpoznawanie separatorów wirtualnych. Zaproponowana metoda ich wykrywania daje wyniki zadowalające, lecz w niektórych warunkach separatory są nieciągłe lub nadmiarowe. Przekłada się to na powstanie błędów na dalszych etapach przetwarzania dokumentu.

Kolejnym aspektem jest wydajność czasowa programu, którą należy określić jako *wystarczającą na potrzeby badań laboratoryjnych*. Wykorzystanie biblioteki *PDFBox* pozwoliło na specjalizację rasteryzacji, niemniej zostało to okupione znacznym narzutem czasowym. Optymalizację wydajnościową należy wskazać jako drugi punkt ewentualnych dalszych badań.

Bibliografia

[1] Adobe System Inc. Adobe digital editions home. Online. http://www.adobe.com/products/digital-editions.html (dostęp: grudzień 2013).

[2] Adobe Systems Inc. Indesing cc. Online. http://www.adobe.com/pl/products/indesign.html (dostęp: grudzień 2013).

[3] Antonacopoulos, A.; Ritchins, R. Flexbile page segmentation using the background. *Pattern Recognition, 1994. Vol. 2 - Conference B: Computer Vision Image Processing., Proceedings of the 12th IAPR International. Conference on 2* (1994), 339–344.

[4] Antonacopoulos, A.; Ritchins, R. Representation and classification of complex-shpaed printed regions using white tiles. *Document Analysis and Recognition, 1995., Proceedings of the Third International Conference on 2* (1995), 1132 – 1135.

[5] Artifex Software, Inc. Mupdf. Online. http://www.mupdf.com (dostęp: grudzień 2013).

[6] Beretta, R.; Laura, L. Performance evaluation of algorithms for newspaper article identification. *Document Analysis and Recognition (ICDAR), 2011 International Conference on* (2011), 394–398.

[7] Bloechle, J.-L. Physical and logical structure recognition of pdf documents. Rozprawa doktorska, 2010 Fribourg.

[8] Bloechle, J-.L.; Rigamonti, M. I. R. Ocd dolores - recovering logical structures for dummies. *Document Analysis Systems (DAS), 2012 10th IAPR International Workshop on* (2012), 245–249.

[9] Fan, J. Text segmentation of consumer magazines in pdf format. *Document Analysis and Recognition (ICDAR), 2011 International Conference on* (2011), 794–798.

[10] Foundation, T. A. S. Apache pdfbox - a java pdf library. Online. http://pdfbox.apache.org/index.html (dostęp: grudzień 2013).

[11] Gatos, B.; Mantazris, S. Inregrated algorithms for newspaper page decomposition and artcile tracking. *Document Analysis and Recognition, 1999. ICDAR '99. Proceedings of the Fifth International Conference on* (1999), 559–562.

[12] Ghostscript an interpreter for the postscript language and for pdf. Online. http://www.ghostscript.com/ (dostęp: grudzień 2013).

[13] Hassan, T.; Baumgartner, R. Table recognition and understanding from pdf files. *Document Analysis and Recognition, 2007. ICDAR 2007. Ninth International Conference on 2* (2007), 1143–1147.

[14] History of ghostscript versions 1.n. Online. http://ghostscript.com/doc/current/History1.htm#Version1.0 (dostęp: grudzień 2013).

[15] Inc., A. S. Document managment - portable document format pdf 1.7. Online. http://www.adobe.com/devnet/pdf/pdf_reference.html (dostęp: grudzień 2013).

[16] Iwanowski, M. *Metody morfologiczne w przetwarzaniu obrazów cyfrowych*. Akademicka Oficyna Wydawnicza EXIT, 2009.

[17] Jełowicki, J. Zajęcia z podstaw informatyki. Online. http://karnet.up.wroc.pl/ jasj/cwiczenia/semi5.html (dostęp: grudzień 2013).

[18] JetBrains. Intellij idea. Online. http://www.jetbrains.com/idea/(dostęp: grudzień 2013).

[19] Librarian of Congress. Starodruk - skan. Online. http://chroniclingamerica.loc.gov/lccn/sn84022355/1836-01-05/ed-1/seq-1.pdf (dostęp: grudzień 2013).

[20] Nashruddin, A. Hit-or-miss transform in opencv. Online. https://opencv-code.com/tutorials/hit-or-miss-transform-in-opencv/ (dostęp: grudzień 2013).

[21] Nieniewski, M. *Morfologia matematyczna w przetwarzaniu obrazów.* Akademicka Oficyna Wydawnicza EXIT, 2005.

[22] Opencv. Online. http://opencv.org/ (dostęp: grudzień 2013).

[23] Page plus. Online. http://www.serif.com/pageplus/ (dostęp: grudzień 2013).

[24] Postscript language tutorial and cookbook. Online. http://www-cdf.fnal.gov/offline/PostScript/BLUEBOOK.PDF (dostęp: grudzień 2013).

[25] Quark Software Inc. Quark. Online. http://www.quark.com/Products/QuarkXPress/(dostęp: grudzień 2013).

[26] Sribus open source desktop publishing. Online. http://www.scribus.net/canvas/Scribus(dostęp: grudzień 2013).

[27] Steć, P. Filtracja obrazów rastrowych. Online. http://www.uz.zgora.pl/ pstec/files/filtracja.pdf (dostęp: grudzień 2013).

[28] Szabłowski, M. Analiza porównawcza wybranych bibliotek do przetwarzania obrazów. Praca magisterska, 2006 Politechnika Warszawska.

[29] Szczurek, G. Zastosowanie metod morfologii matematycznej do detekcji i dekompozycji obrazów. *Telekomunikacja i Techniki informacyjne* (1-2/2003).

[30] Tadeusiewicz, R.; Korohoda, P. *Komputerowa analiza i przetwarzanie obrazów*. Wydawnictwo Fundacji Postępu Telekomunikacji, 1997.

[31] Wikiepdia. Page description language. Online. http://en.wikipedia.org/wiki/Page_description_language (dostęp: grudzień 2013).

[32] Wikiepdia. Postscript. Online. http://en.wikipedia.org/wiki/PostScript (dostęp: grudzień 2013).