وزارت علوم، تحقیقات و فناوری
پژوهشگاه اطلاعات و مدارک علمی ایران

فصلنامه علوم و فناوری اطلاعات

دوره ۲۱شمارهء ۳ (بهار ۱۳۸۵) صفحه: ۱-۳۴

تشخیص خودکار صفحات فهرست با توجه به الگوی آن‌ها در پایان‌نامه‌های فارسی و لاتین

۱) اسماعیل فرامرزی

چکیده: صفحات فهرست در هر نوع مدرک (کتاب، مجله، پایان‌نامه، ...)، به نحو مختصر و فشرده، ساختار منطقی آن مدرک را بیان می‌کنند و به کمک آن‌ها می‌توان به راحتی ساختار مدرک را مشاهده نمود و مستقیماً به مطالب مورد نظر دست یافت. در این مقاله برای اولین بار روشی به منظور شناسایی خودکار صفحات فهرست در پایان‌نامه‌های فارسی، عربی و لاتین ارائه می‌شود. در این روش، شناسایی صفحات فهرست با توجه به الگوی آن‌ها بدون استفاده از عملیات بازشناسی متن (اُسی‌آر) و تنها با به‌کارگیری فنون پردازش تصویر، مد نظر بوده. با این روش می‌توان صفحات فهرست را صرف نظر از نوع زبان و ترازبندی[۱] (راست به چپ یا چپ به راست بودن) متن آن‌ها، شناسایی کرد و به دلیل عدم استفاده از اُسی‌آر، تابع کیفیت متن مدرک اسکن‌شده نیست. روش مذکور بر روی دسته‌ای از پایان‌نامه‌های فارسی، عربی و لاتین موجود در پایگاه اطلاعاتی پژوهشگاه اطلاعات و مدارک علمی ایران مورد آزمایش قرار گرفت و دقت ۹۹/۷ درصد در بازشناسی صحیح حاصل گردید.

کلیدواژه‌ها: تحلیل تصویر مدارک (دی‌آی‌ای)[۲]، تحلیل پیکربندی صفحات[۳]، تحلیل ساختاری[۴] مدارک، تحلیل منطقی[۵] مدارک، درک تصویر مدارک[۶]، شناسایی صفحات فهرست[۷]، پردازش تصویر[۸]، بازشناسی نوری حروف (اُسی‌آر) [۹]، شناسایی الگو[۱۰]، کتابخانهٔ دیجیتال.

۱. مقدمه[۱۱]

در سه دههٔ گذشته همگام با ظهور و پیشرفت رایانه‌ها، فنون پردازش سیگنال‌های دیجیتال اهمیت خاصی یافته‌اند. تصاویر دیجیتال که یا مستقیماً توسط دستگاه‌های

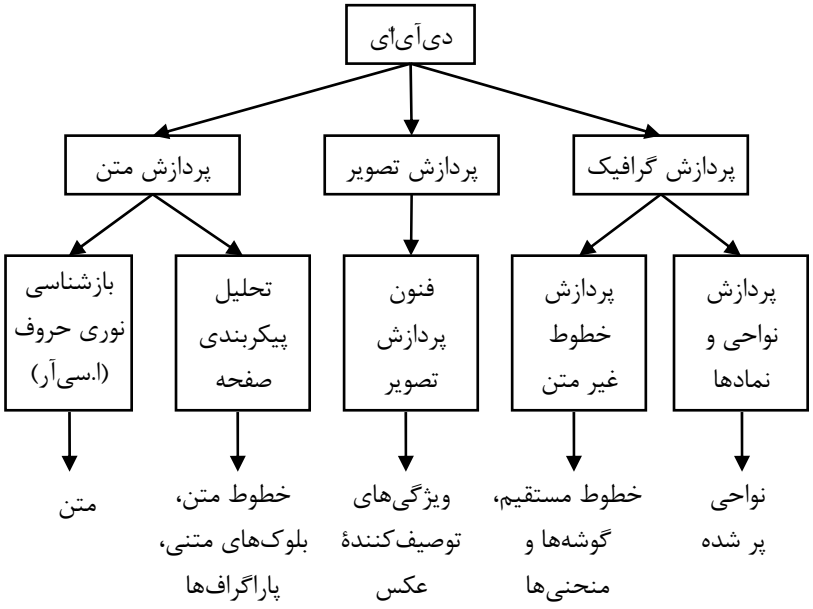
تصویربرداری دیجیتال اخذ می‌شوند یا بۀوسیلۀ دستگاه‌هایی نظیر اسکنر، دوربین و فکس از فرم چاپی به دیجیتالی تبدیل می‌گردند، دستۀ بسیار مهمی از سیگنال‌های دیجیتالی (دو بعدی) را تشکیل می‌دهند. این تصاویر با استفاده از فنون پردازش تصویر، شناسایی الگو، بازشناسی نوری حروف (اُسی‌آر) و تحلیل تصویر مدارک (دی‌آی‌ای) مورد تجزیه و تحلیل قرار می‌گیرند.

امروزه حجم زیادی از مدارک کاغذی موجود، توسط اسکنرها یا دوربین‌ها به تصاویر دیجیتالی تبدیل می‌شوند. ذخیره‌سازی، بازیابی و مدیریت کارآمد این آرشیوهای تصویری، در بسیاری از کاربردها نظیر اتوماسیون اداری و کتابخانه‌های دیجیتالی اهمیت فراوانی دارد. در نتیجه، دستیابی به الگوریتم‌های مؤثر به‌منظور تحلیل تصویر مدارک، یک نیاز اساسی به حساب می‌آید (فرامرزی، ۱۳۸۳، ۱۳۸۴ب).

مبحث «تحلیل تصویر مدارک» (دی‌آی‌ای) از جمله شاخه‌های فعال در زمینه‌های شناسایی الگو و پردازش تصاویر می‌باشد و مشتمل بر کلیۀ مراحل پردازشی است که محتویات یک مدرک اسکن یا فکس شده را به فرم الکترونیکی مناسب، کد می‌نماید. این کدکردن چندین شکل دارد. یک توصیف قابل ویرایش، یک نمایش فشرده که تصویر مدرک از آن قابل بازیابی باشد، یا یک توصیف معناشناختی سطح بالا که به منظور پاسخگویی به پرس‌وجوها[۱۲] به‌کار می‌رود (Srihari, et al., 1992). فنون دی‌آی‌ای، اجزای مختلف ساختاری مدرک یعنی قسمت‌های متنی (پاراگراف‌ها، کلمات، حروف، ...)، قسمت‌های گرافیکی (خطوط، نمادها، نمودارها، ...) و عکس‌ها (تصاویر موجود در متن) را از یکدیگر تفکیک می‌کنند و پردازش مناسب را بر روی هر دسته از اجزا اعمال می‌نمایند. شکل ۱ ساختار سلسله مراتبی «دی‌آی‌ای» را به نمایش می‌گذارد (O'Gorman, 1995, p. 2).

همانگونه که شکل ۱ نشان می‌دهد، در مبحث «دی‌آی‌ای» سه دسته فنون پردازشی زیر مد نظر است:

الف. پردازش متن: این فنون با اجزای متنی تصویر مدرک سر و کار دارند. برخی از فنون این گروه عبارت‌اند از: تعیین و تصحیح زاویۀ کجی مدرک، بازشناسی نوری حروف (اُسی‌آر) و یافتن ستون‌ها، پاراگراف‌ها، خطوط متن و کلمات.



شکل ۱ ساختار سلسله‌مراتبی بخش‌های مختلف مبحث تحلیل تصویر (دی‌آی‌دی)

ب. پردازش گرافیک[۱۳]: فنون این دسته، با خطوط و نمادها سر و کار دارند و مواردی نظیر نمودارهای خطی، خطوط مستقیم مرزبندی بین قسمت‌های متنی، آرم شرکت‌ها و غیره را شامل می‌شوند. از آنجا که بسیاری از اجزای گرافیکی از خطوط تشکیل شده‌اند، این دسته از فنون پردازشی مشتـمل بر نازک‌سازی خطوط، برازش خطی[۱۴]، آشکارسازی گوشه‌ها و منحنی‌ها می‌باشد.

ج. پردازش عکس[۱۵]: عکس‌ها سومین جزء اصلی مدارک هستند که البته بجز تشخیص موقعیت مکانی آن‌ها در تصویر مدرک، تحلیل‌های دیگر معمولاً به دیگر فنون پردازش تصویر و بینایی ماشین مربوط می‌شود. مهم‌ترین این پردازش‌ها، فشرده‌سازی[۱۶] است (به Gonzalez & Woods, 2002, pp. 409-518 مراجعه شود).

دو دسته فنون پردازشی وجود دارند که بر روی متن موجود در مدارک اعمال می‌شوند (O'Gorman, 1995, p. 161):

- بازشناسی نوری حروف (اُسی‌آر)،

- تحلیل پیکربندی صفحات.

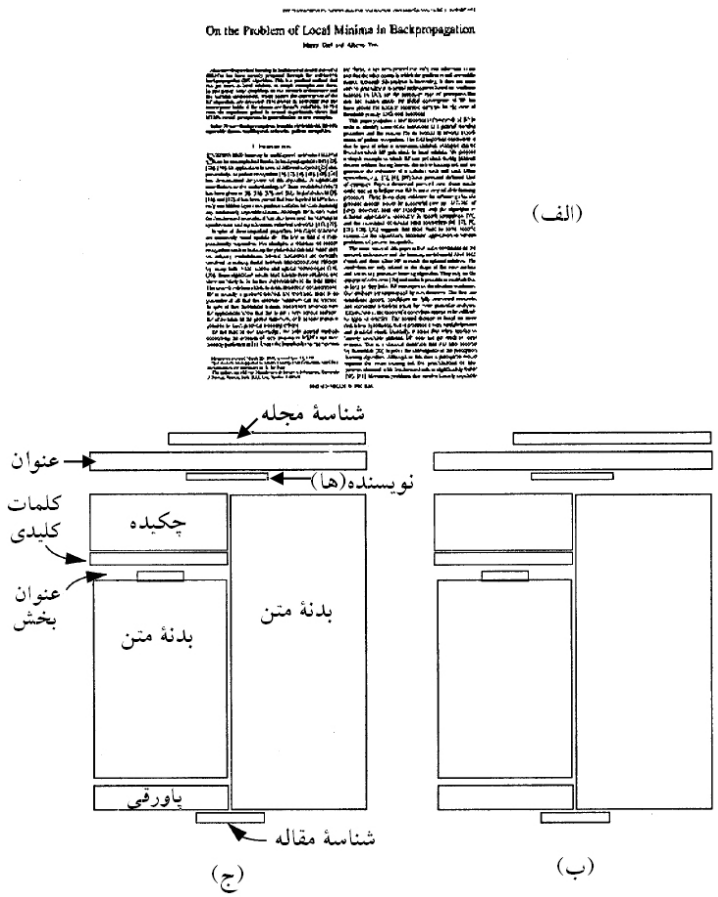
این دو عملیات را می‌توان بصورت جداگانه بر روی متن اعمال کرد، یا این‌که نتایج حاصل از یکی، در جهت تصحیح یا کمک به دیگری عمل نماید. فنون اُسی‌آر، حروف و کلمات را در تصویر مدرک شناسایی می‌نمایند. ورودی یک سیستم اُسی‌آر، یک فایل تصویری حاوی تصویر مدرک می‌باشد و خروجی آن، یک فایل متنی است[۱۷]. فنون تحلیل پیکربندی صفحات خود به دو دسته تقسیم می‌گردند:

- تحلیل پیکربندی ساختاری (یا فیزیکی یا هندسی)،
- تحلیل پیکربندی منطقی (یا نحوی[۱۸] یا عملکردی[۱۹]).

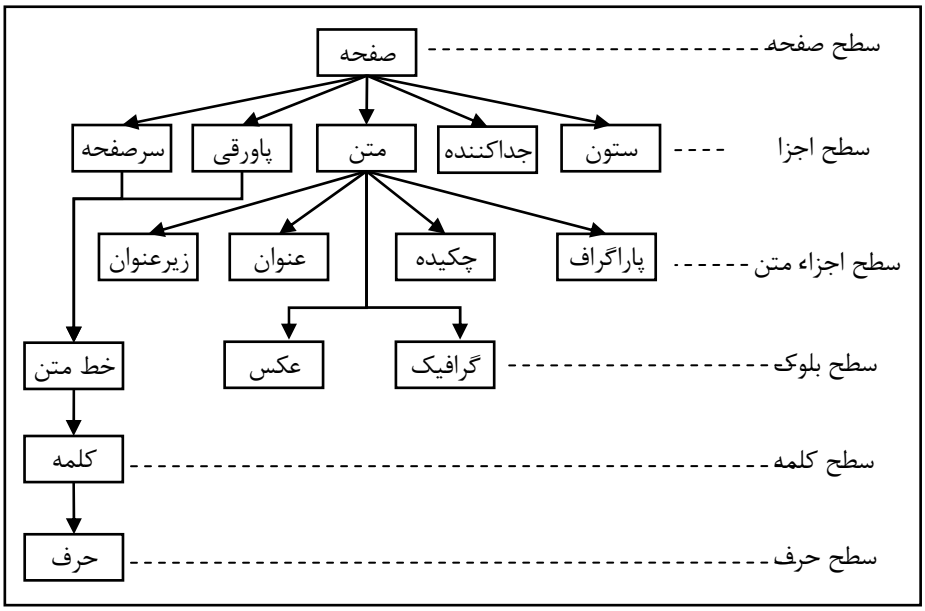
تحلیل پیکربندی ساختاری به‌منظور قطعه‌بندی[۲۰] متن مدرک به گروه‌های مختلف، مورد استفاده قرار می‌گیرد. بسته به فرمت مدرک، این گروه‌ها ممکن است کلمات مجزا، خطوط متن، و بلوک‌های ساختاری نظیر پاراگراف‌ها، ستون‌ها، هر یک از بخش‌های صفحهٔ فهرست، و غیره باشند. تحلیل پیکربندی ساختاری می‌تواند به‌صورت الگوریتم‌های بالا به پایین[۲۱] یا پایین به بالا[۲۲] پیاده‌سازی گردد. در تحلیل بالا به پایین، صفحه از اجزای بزرگ به زیر جزءهای کوچک‌تر قطعه‌بندی می‌شود. به عنوان مثال صفحه ممکن است به یک یا چند بلوک ستونی متنی تقسیم گردد؛ سپس هر ستون نیز امکان دارد به بلوک‌های پاراگرافی تقسیم‌بندی شود و هر پاراگراف نیز به نوبهٔ خود به خطوط متن، مجزا گردد و این روند ادامه یابد. در مورد تحلیل پایین به بالا، اجزای پیوسته به ترتیب به کاراکترها، کلمات، خطوط متن و اجزای بزرگ‌تر گسترش می‌یابند.

تحلیل پیکربندی منطقی، از اطلاعات محدودبه دامنه[۲۳] (شامل قوانین پیکربندی یک صفحهٔ خاص) استفاده می‌کند تا بلوک‌های ساختاری مدرک را برچسب‌گذاری[۲۴] نماید و بدین ترتیب نوعی نشانه‌گذاری[۲۵] از عملکرد بلوک‌ها انجام می‌دهد. این نوع نشانه‌گذاری عملکردی ممکن است مستلزم به‌کارگیری فنون تقسیم[۲۶] یا ادغام[۲۷] بلوک‌های ساختاری باشد. به عنوان مثال با اِعمال فنون تحلیل پیکربندی منطقی به صفحات یک مقاله، می‌توان آن را به بخش‌هایی چون عنوان مقاله، مشخصات نویسندگان مدرک، کلمات کلیدی، بخش‌ها و زیربخش‌های متن اصلی، زیرنویس‌ها و غیره گروه‌بندی نمود[۲۸].

شکل ۲ مثالی از نتیجهٔ حاصل از اعمال تحلیل‌های پیکربندی ساختاری و عملکردی را برای صفحهٔ اول یک مقالهٔ فنی نشان می‌دهد (O'Gorman, 1995, p. 166). تحلیل پیکربندی ساختاری، بلوک‌هایی را نشان می‌دهد که بر پایهٔ فواصل موجود در تصویر اصلی مقاله، قطعه‌بندی گردیده‌اند. عملیات برچسب‌گذاری در مرحلهٔ تحلیل پیکربندی منطقی، با توجه به اطلاعات قبلی موجود دربارهٔ قوانین صفحه‌بندی این مقالهٔ ژورنال انجام شده است. شکل ۳ نیز سلسله‌مراتب ساختار منطقی یک مدرک را به نمایش می‌گذارد (Wieser, & Pinz, 1993).



شکل ۲ الف. صفحهٔ اول یک مقالهٔ فنی ب. پیکربندی ساختاری ج. پیکربندی منطقی



شکل ۳ سلسله‌مراتب ساختار منطقی یک مدرک

از آنجا که فرآیند تحلیل پیکربندی ساختاری و منطقی، برعکس فرایند فرمت‌بندی مدارک می‌باشد، منطقی است که برای ایجاد و توصیف محتوای مدارک به دنبال استانداردی باشیم که بتوان از آن به منظور انجام تحلیل پیکربندی و سنجش نتایج استفاده نمود. در گذشته دو استاندارد «اُدی‌اُی»[۲۹] (ITU-T, Website) و «اس‌جی‌اُم‌اُل»[۳۰] (W3C: …, Overview of SGML resources; W3C:… The Extensible Stylesheet Language family (XSL) برای توصیف ساختار و محتوای متن مورد استفاده قرار می‌گرفتند که امروزه جای خود را به استانداردهای «اچ‌تی‌ام‌ال»[۳۱]، «سی‌اس‌اس»[۳۲]، «ایکس‌ام‌ال»[۳۳]، «ایکس‌اس‌ال»[۳۴] و «اُدی‌اف»[۳۵] (OASIS: Website) داده‌اند[۳۶].

همگام با پیشرفت نظری در مباحث «تحلیل تصویر مدارک» و «درک تصویر مدارک»، کتابخانه‌های دیجیتالی پا به عرصهٔ ظهور گذاشته‌اند که مدیریت هر چه بهتر مدارک کاغذی را در فرم الکترونیکی به منظور تسهیل در عملیات نمایه‌سازی[۳۷]، نمایش، چاپ و استخراج قسمت‌های مورد نظر، تحقق بخشند. برای کتابخانه‌های

دیجیتالی، تشخیص صفحات فهرست در بین سایر صفحات اسکن‌شدهٔ مدارک دارای اهمیت فراوانی است؛ چرا که در هر مدرک (از جمله مجله، کتاب، پایان‌نامه)، این صفحات به نحو مناسبی ساختار منطقی مدرک را بیان می‌کنند و توسط آن‌ها می‌توان مطالب مطروحه در مدرک و نحوهٔ تنظیم آن‌ها را به سرعت مشاهده نمود. همچنین این امکان وجود دارد که با برقراری پیوند (لینک[۳۸]) میان عناوین بخش‌ها در صفحات فهرست و صفحات اصلی (ایجاد کتاب‌نشان[۳۹])، سرعت دسترسی به مطالب مندرج در مدارک را به نحو چشمگیری افزایش داد؛ به نحوی که کاربران خواهند توانست با کلیک بر روی هر یک از عناوین صفحات فهرست، بدون نیاز به مرور صفحات قبل مستقیماً صفحهٔ مربوطه در مدرک را مشاهده نمایند. برخی از پایگاه‌های دادهٔ دیجیتالی نیز از اطلاعات صفحات فهرست به‌منظور ساختارسنجی و نمایه‌سازی مدارک بهره می‌جویند.

با مروری بر منابع علمی موجود مشاهده می‌گردد که بیش‌تر تحقیقات انجام‌شده، درک سطح بالای صفحات فهرست را مد نظر قرار داده‌اند تا بدینوسیله، اطلاعات ساختاری آن‌ها را استخراج نمایند و به فرم اَبَرساختاری[۴۰] نظیر «اچ‌تی‌ام‌ال» یا «ایکس‌ام‌ال» نمایش دهند. برای این منظور فرض گردیده است که صفحات فهرست یا قبلاً قطعه‌بندی شده‌اند، یا برخی فنون بازشناسی حروف/نمادها[۴۱] برای تعیین این صفحات و ساختار زیربنایی آن‌ها مورد استفاده قرار می‌گیرند (Mandal, et al., 2003). به طور کلی، استفادهٔ صرف از اُسی‌آر برای تشخیص صفحات فهرست بدون توجه به اطلاعات ساختاری، به دلایل زیر مناسب نمی‌باشد:

الف. اگر قرار باشد فنون بازشناسی حروف/نمادها که جزئی از عملیات اُسی‌آر می‌باشند برای یک صفحهٔ حاوی مطالب ناهمسان و تفکیک‌نشده (متن همراه با فرمول‌های ریاضی[۴۲]) اعمال گردند، حجم محاسبات زیاد می‌باشد، و نتیجهٔ کار رضایت‌بخش نخواهد بود (Mandal, et al., 2003).

ب. فنون اُسی‌آر برای زبان‌های فارسی و عربی که نگارش پیوسته[۴۳] دارند، در حد قابل قبول توسعه نیافته‌اند.

ج. در مواردی که مجموعهٔ آزمایشی مورد استفاده حاوی مدارک چندزبانه اعم از لاتین و غیر لاتین است، یک سامانهٔ اُسی‌آر چندزبانه با قابلیت تشخیص خودکار زبان متن صفحه، مورد نیاز می‌باشد. در حال حاضر چنین سامانه‌ای با کارآیی قابل قبول (چه

در بعد تحقیقاتی و چه در بعد تجاری) وجود ندارد. از طرف دیگر به جز عامل چپ به راست یا راست به چپ بودن متن صفحهٔ فهرست که تابع زبان نگارش مدرک است، بیش‌تر اطلاعات ساختاری، از زبان مدرک تأثیر نمی‌پذیرند و از این رو یک الگوریتم شناسایی صفحات فهرست که مبتنی بر اطلاعات ساختاری است می‌تواند صفحات فهرست را - قطع نظر از زبان نگارش این صفحات - تشخیص دهد.

د. پایین بودن کیفیت مدارک قدیمی اسکن‌شده، بازشناسی صحیح در عملیات اُسی‌آر را با مشکل مواجه می‌سازد. این در حالی است که اطلاعات ساختاری در این مدارک به مقدار زیادی محفوظ می‌ماند.

در این مقاله برای اولین بار روشی به منظور شناسایی خودکار صفحات فهرست (اعم از فهرست مطالب، فهرست شکل‌ها، فهرست جداول، فهرست اختصارات، و ...) با استفاده از فنون پردازش تصویر و تحلیل تصاویر مدارک و بدون به‌کارگیری فنون بازشناسی متن (اُسی‌آر) در پایان‌نامه‌های فارسی، عربی و لاتین ارائه می‌گردد. در این روش تنها از اطلاعات ساختاری که اکثر نویسندگان در مورد صفحات فهرست رعایت می‌کنند، استفاده گردیده است. به جهت استفادهٔ صرف از اطلاعات ساختاری و عدم به‌کارگیری اُسی‌آر، این روش تابع نوع زبان مدرک، ترازبندی متن و کیفیت متن صفحات اسکن‌شده نیست. روش مذکور بر روی دسته‌ای از پایان‌نامه‌های موجود در پایگاه اطلاعاتی پژوهشگاه اطلاعات و مدارک علمی ایران مورد آزمایش قرار گرفته است.

نحوهٔ تنظیم مطالب در این مقاله بدین شرح است: در بخش ۲ کارهای محققان دیگر در زمینهٔ استخراج و به‌کارگیری اطلاعات صفحات فهرست مورد اشاره قرار می‌گیرد. در بخش ۳ ویژگی‌های خاص صفحات فهرست در پایان‌نامه‌ها معرفی می‌شوند. بخش ۴ به تشریح روش معرفی‌شده در این مقاله به منظور شناسایی خودکار صفحات فهرست در پایان‌نامه‌ها به کمک اطلاعات ساختاری می‌پردازد. بالاخره در بخش ۵ نتایج حاصل از پیاده‌سازی روش پیشنهادی را بررسی خواهیم نمود.

۲. کارهای دیگران

«اگورمن» و «استوری و همکارانش» در پروژه‌ای با عنوان Right Pages Electronic Library Systems، روشی به منظور استخراج ساختار صفحات فهرست با به‌کارگیری

فنون اُسی‌آر پیشنهاد نموده‌اند. با استفاده از شیوه‌ای به نام «داکستروم»[۴۴]، ابتدا بلوک‌ها استخراج می‌شوند و سپس مدارک به کمک یک مدل ازپیش‌معین[۴۵] (برگرفته از صفحات فهرست ژورنال‌های مختلف)، نمایه‌سازی می‌گردند. مدل مذکور به‌صورت دستی (غیر خودکار) به سامانه داده می‌شود. همچنین از اطلاعات پیوندی استفاده می‌شود تا کاربر براحتی بتواند با کلیک کردن بر روی عناوین بخش فهرست، به صفحات مربوطه مراجعه نماید (O'Gorman, 1992; Story, et al., 1992).

«تاکاسو، ساتو و کاتسورا» بر مبنای قطعه‌بندی تصاویر به بلوک‌ها و تحلیل نحوی محتوای آن‌ها، سامانه‌ای به نام CyberMagazine را برای تحلیل صفحات فهرست ارائه داده‌اند که در آن، تلفیقی از طبقه‌بندی با استفاده از درخت تصمیم‌گیری، و تحلیل نحوی با کمک گرامر ماتریسی به کار گرفته می‌شود. ورودی سامانه، فایل‌های متنی صفحات فهرست می‌باشد (Satoh, Takasu, & Katsura, 1995; Takasu, Satoh, & Atsura, 1994; 1995).

«لین، نیوا و ناریتا» شیوه‌ای به منظور تحلیل ساختار منطقی مدارک برای تبدیل تصاویر مدارک و اطلاعات صفحات فهرست به فرم الکترونیکی مناسب، ابداع نموده‌اند. در این روش ابتدا صفحات فهرست با کمک اُسی‌آر خوانده می‌شوند و اطلاعات متنی آن‌ها تحلیل می‌گردد تا ساختار منطقی کل مدرک به دست آید. سپس صفحات متنی بدنهٔ اصلی مدرک، وارد سامانه می‌شوند و اجزای مختلف شامل شمارهٔ صفحات، پاورقی‌ها و سرصفحه‌ها، عناوین بخش‌ها و متن عادی، تحلیل می‌گردند. در نهایت اطلاعات مربوط به عناوین بخش‌ها که از صفحات متنی بدنهٔ اصلی مدرک استخراج شده بودند، با اطلاعات به‌دست‌آمده از صفحات فهرست تطبیق داده می‌شوند تا دقت شناسایی عناوین، بهبود یابد (Lin, Niwa, & Narita, 1997).

«بلاید، پیرون و والورد» با الهام از شیوهٔ تگ‌زنی[۴۶] سیگنال صوتی، الگوریتمی با عنوان «پی‌اُاُس»[۴۷] برای شناسایی خودکار صفحات فهرست معرفی کرده‌اند. در این روش از یک مدل ازپیش‌معین برگرفته از نظام[۴۸] موجود در ساختار و محتوای مدرک استفاده شده است. شیوهٔ تگ‌زنی پی‌اُاُس مستقیماً و بدون انجام هرگونه پیش‌پردازش، روی فایل‌های متنی تولیدشده بوسیلهٔ عملیات اُسی‌آر به‌کار برده می‌شود (Belaid, Pierron, & Valverde, 2000).

«تسوروکا و دیگران»، روشی به منظور تحلیل ساختار مبتنی بر تصویر[۴۹] صفحات فهرست کتاب‌ها و تبدیل آن‌ها به فرمت «ایکس‌ام‌ال» پیشنهاد نموده است. در این مقاله چنین فرض شده که اجزای ساختاری صفحات فهرست نظیر عناوین فصل‌ها، بخش‌ها و غیره با مقدار تورفتگی[۵۰] و اندازهٔ فونت حروف آن‌ها از یکدیگر متمایز می‌شوند. این اطلاعات برای دسته‌بندی خطوط متن به گروه‌های مختلف، به‌کار گرفته شده‌اند. برای عملیات تبدیل به ایکس‌ام‌ال، یک درخت ساختاری ازپیش‌آماده‌شده، به گروه‌ها ارتباط داده می‌شود. اما کاربرد این شیوه تنها به کتاب‌هایی که بخش‌های مختلف صفحات فهرست در آن‌ها دارای تورفتگی و اندازهٔ فونت متفاوت هستند، محدود می‌گردد (Tsuruoka, et al., 2001).

«لی برجوس، امپتوز و سوفی بن صافی» به توصیف یک مدل آماری برای یک سامانهٔ درک مدارک پرداخته‌اند که در آن از ویژگی‌های متنی و قواعد پیکربندی مدارک بهره‌گیری شده است. در این مدل، از شیوهٔ آرام‌سازی آماری[۵۱] برای یافتن ساختار سلسله‌مراتبی پیکربندی منطقی استفاده شده است که کاملاً با مدل آموزشی سازگار می‌باشد. از این مدل برای خواندن متن اُسی‌آر شدهٔ صفحات فهرست به‌منظور نمایه‌سازی مدارک استفاده شده است (Le Bourgeois, Emptoz, & Souafi Bensafi, 2001).

شناسایی خودکار صفحات فهرست با استفاده از اطلاعات ساختاری و بدون به‌کارگیری اُسی‌آر، برای نخستین و تنها بار توسط «مندل و همکارانش» پیشنهاد گردیده است. در این مقاله با ملاحظهٔ ۸۲ صفحهٔ فهرست اسکن‌شده از کتاب‌ها، مجلات علمی و گزارش‌ها، مدلی برای صفحات فهرست استخراج شده است. بر اساس این مدل، دو نوع صفحهٔ فهرست در نظر گرفته می‌شود: در نوع اول شمارهٔ صفحهٔ متناظر با هر عنوان فهرست، ترازبندی راست‌چین[۵۲] دارد (TOC-I) و در نوع دوم، شمارهٔ صفحات تنها به فاصلهٔ کمی از عنوان و بدون ترازبندی قرار می‌گیرند (TOC-II). چنانچه سمت راست‌ترین کلمات خطوط یک بلوک متن، ترازبندی راست‌چین داشته و همچنین هیستوگرام عمودی بلوک متن، یک قلهٔ باریک در سمت راست داشته باشد، آن بلوک متن به عنوان TOC-I در نظر گرفته خواهد شد. از طرفی اگر سمت راست‌ترین کلمات، تراز بندی راست‌چین نداشته باشند اما تغییر در تعداد حروف آن‌ها، به صورت صعودی و

به آهستگی صورت پذیرد، بلوک متن به عنوان TOC-II شناخته خواهد شد. سرانجام نیز از یک درخت تصمیم‌گیری[۵۳] برای برچسب‌گذاری خطوط فهرست به بخش‌های شمارهٔ عنوان، عنوان و شمارهٔ صفحه استفاده گردیده است (Mandal, et al., 2003).

«لین» روشی به منظور تقسیم محتوای یک ژورنال به مقالات موجود در آن ارائه داده است. .در این روش از یک الگوریتم متن‌کاوی[۵۴] برای آشکارسازی عبارت‌های منطبق بین صفحات فهرست و صفحات محتوای هر یک از مقالات استفاده شده است (Lin, 2003).

روشی برای تحلیل صفحات فهرست به‌منظور درک ساختار منطقی کتاب‌های منتشره به زبان چینی توسط «هی، دینگ و پنگ» ارائه شده است که در آن از ترکیب قوانین هندسی (تورفتگی خطوط) و قوانین معناشناختی (توالی معمول متن فصل‌ها و بخش‌های مدرک) برای استخراج ساختار منطقی سلسله‌مراتبی کتب به کار گرفته می‌شود. ابتدا سطور متن، استخراج و ناحیه‌بندی می‌شوند. سپس عملیاتی بر روی سطور متن در هر ناحیه صورت می‌گیرد که شامل تقطیع سطور متن، یکی کردن سطور مربوط به هر عنوان فهرست، و تعیین سطح هر عنوان با توجه به میزان تورفتگی و شماره عنوان می‌باشد. پس از آن کلیهٔ عناوین فهرست موجود در هر ناحیه، در توالی صحیح یکپارچه می‌گردند. در نهایت اندیس‌های تعیین‌شده برای هر صفحهٔ فهرست به یکدیگر اضافه می‌شوند تا بدین طریق ساختار منطقی سلسله‌مراتبی کل کتاب به دست آید (He, Ding, & Peng, 2003).

«یاکوب و پیرو» روشی برای درک ساختار مدارک و شناسایی صفحات فهرست در مدارک اسکن و اُسی‌آر شده معرفی کرده‌اند که از سه شیوهٔ مختلف به منظور ارزیابی صفحات فهرست و تعیین موقعیت مقالات استفاده شده است. این سه شیوه عبارت‌اند از: تطبیق‌دهی[۵۵] عناوین، تطبیق‌دهی کلیدواژه‌هایی که احتمال وجود آن‌ها در صفحهٔ فهرست وجود دارد، و تعیین صفحهٔ ابتدایی هر مقاله با کمک شمارهٔ ذکرشدهٔ آن در صفحهٔ فهرست. بر اساس هر یک از فنون فوق، صفحاتی به عنوان کاندیدهای صفحهٔ فهرست بودن یا صفحهٔ شروع مقاله بودن مشخص می‌گردند. سپس با استفاده از یک الگوریتم ترکیب، نمرهٔ هر صفحهٔ کاندیدا تعیین می‌شود و از میان آن‌ها بهترین کاندیداها مشخص می‌گردند (Yacoub & Peiro, 2005).

الگوریتمی به منظور آشکارسازی و تحلیل صفحات فهرست توسط «لین و زیانگ» برای پروژه‌ای با عنوان DCRM[۵۶] تدوین شده است. هدف از این پروژه تبدیل کتاب‌ها، مجلات و نشریات منتشر شده توسط «انتشارات امآی‌تی»[۵۷] از فرم کاغذی به فرم الکترونیکی می‌باشد. مبنای این الگوریتم روش «هم‌پیوندی محتوایی[۵۸]» است و مد نظر طراحان آن این بوده که الگوریتم آن‌ها بتواند صفحات فهرست موجود در انواع متنوع مدارک را بدون نیاز به هرگونه مدل‌سازی یا آموزش، شناسایی و تحلیل نماید. ابتدا چند صفحه از مدرک به عنوان کاندیدا برای صفحات فهرست انتخاب می‌شوند و سپس با کمک دو الگوریتم متن‌کاوی و تطبیق‌دهی شمارهٔ صفحات، محتوای هر یک از این صفحات با صفحات اصلی مرتبط می‌گردند. در نهایت صفحات فهرست واقعی از طریق یک سازوکار نمره‌دهی، از بین صفحات فهرست کاندیدا تعیین می‌شوند. به منظور برقراری پیوند بین هر سطر از صفحهٔ فهرست به صفحهٔ متناظر آن در مدرک، از یک الگوریتم برای تشخیص مکان درج شماره صفحهٔ هر یک از صفحات اصلی استفاده می‌شود. کلمات، گروه‌های کلمات و بلوک‌های هر سطر فهرست با به‌کارگیری یک روش کلاسه‌بندی پایین به بالا[۵۹] مشخص خواهند شد. در پایان از اطلاعات پیکربندی به همراه یک الگوریتم تگ‌زنی برای برچسب‌زنی عملکردی صفحات فهرست استفاده می‌شود (Lin & Xiong, 2006).

۳. مدل صفحات فهرست در پایان‌نامه‌ها

برای شناسایی صفحات فهرست در پایان‌نامه‌ها ابتدا لازم است مدلی برای آن‌ها در نظر بگیریم. همانگونه که در شکل‌های ۴ و ۵ دیده می‌شود، صفحات فهرست در پایان‌نامه‌ها معمولاً دارای مشخصات زیر هستند (گرچه در بسیاری از پایان‌نامه‌ها، همهٔ این موارد بطور دقیق رعایت نگردیده‌اند):

الف. صفحات فهرست در پایان‌نامه‌ها ممکن است شامل انواع فهرست مطالب، فهرست جداول، فهرست اشکال، فهرست نمودارها، فهرست اصطلاحات یا فهرست اختصارات باشند. کلیهٔ این صفحات دارای ساختار تصویری مشابه هستند.

ب. صفحات فهرست دارای ساختار جدولی[۶۰] هستند.

شکل ۴ بخشی از یک صفحهٔ فهرست

ج. در هر خط از صفحات فهرست (جز در فهرست اصطلاحات یا فهرست اختصارات)، شمارهٔ یک عنوان پایان‌نامه به همراه نام عنوان در یک طرف، و شمارهٔ صفحهٔ مربوط به آن در طرف دیگر قرار دارد.

د. شمارهٔ صفحات در پایان‌نامه‌های فارسی و عربی ترازبندی چپ‌چین، و در پایان‌نامه‌های لاتین ترازبندی راست‌چین دارند.

ه. بین متن و شمارهٔ صفحات در هر خط ممکن است خط، نقطه‌چین یا خط‌چین وجود داشته باشد. این «خطوط راهنما» به‌منظور اجتناب از خطای دید چشم در مربوط کردن هر عنوان به شمارهٔ صفحهٔ متناظرش به کار گرفته می‌شوند.

و. شمارهٔ صفحات تنها روند افزایشی دارند.

ی. عناوین طولانی در صفحات فهرست ممکن است در دو خط آمده باشند (شکل ۵). این مورد، بخصوص در صفحات فهرست جداول و فهرست اشکال که نام عناوین طولانی است، به کرات دیده می‌شود.

ز. در برخی از صفحات فهرست، شمارهٔ صفحات به صورت یک عدد نیست، بلکه به شکل یک محدودهٔ عددی آمده است (شکل ۵).

البته در روش پیشنهادی، تمامی این اطلاعات به کار گرفته نشده است.

فصل پنجم: رابطهٔ ضروری میان نفس و بدن از نظر لایب‌نیتس و ملاصدرا

۵-۱-۱- رابطهٔ نفس و بدن از دیدگاه لایب‌نیتس..... ۱۶۶

۵-۱-۲- تعریف نفس وبدن در فلسفه لایب‌نیتس..... ۱۶۷-۱۶۸

۵-۱-۳- هماهنگی پیشین بنیاد میان نفس و بدن..... ۱۶۸-۱۷۰

۵-۲-۱- معنای تجرّد نفس از دیدگاه ملاصدرا..... ۱۷۱-۱۷۵

۵-۲-۲- تسلط نفس بر بدن و جنبهٔ فاعلیت آن (در دیدگاه ملاصدرا)..... ۱۷۵-۱۷۶

۵-۳- علم انسان به صور ذهنی و اشیاء خارجی از نظر لایب‌نیتس و صدرالمتألهین..... ۱۷۶-۱۷۸

۵-۴- ضرورت رابطهٔ نفس و بدن و بحث معاد از دیدگاه لایب‌نیتس و ملاصدرا..... ۱۷۸-۱۸۰

شکل ۵ قسمتی از یک صفحهٔ فهرست دارای عناوین چندخطی و شمارهٔ صفحات محدوده‌ای

۴. معرفی روش پیشنهادی

از آنجا که روش معرفی‌شده در مقالهٔ (Mandal, et al., 2003) اولین و تنها روش معرفی‌شده در مقالات برای شناسایی صفحات فهرست به کمک اطلاعات ساختاری بود، ابتدا آن را پیاده‌سازی نمودیم. با آزمایش، مشخص گردید که این روش برای شناسایی صفحات فهرست در پایان‌نامه‌ها بویژه پایان‌نامه‌های فارسی و عربی چندان مناسب نیست؛ دلیل اصلی این امر، عدم رعایت فرمت‌بندی[۶۱] مناسب در پایان‌نامه‌ها است. باید توجه داشت که پایان‌نامه‌ها توسط دانشجویان نگاشته شده‌اند و نه مؤسسات انتشاراتی. در نتیجه قواعد فرمت‌بندی در آن‌ها (بخصوص در پایان‌نامه‌های قدیمی) به دقت رعایت نشده است. به عنوان نمونه در بسیاری از موارد، شمارهٔ صفحات در صفحات فهرست، فاقد ترازبندی چپ‌چین (در پایان‌نامه‌های فارسی و عربی) یا راست‌چین (در پایان‌نامه‌های لاتین) می‌باشند. همچنین همان‌طور که در بخش ۳ اشاره شده، در برخی

پایان‌نامه‌ها، شمارهٔ صفحات به صورت یک محدوده داده شده است. این دو مورد سبب می‌شود که هیستوگرام عمودی صفحه، یک قلهٔ باریک به ازای شمارهٔ صفحات نداشته باشد. در ادامهٔ این بخش، مرحله به مرحله روش پیشنهادی را تشریح خواهیم نمود.

۴-۱. عملیات پاکسازی صفحه[۶۲]

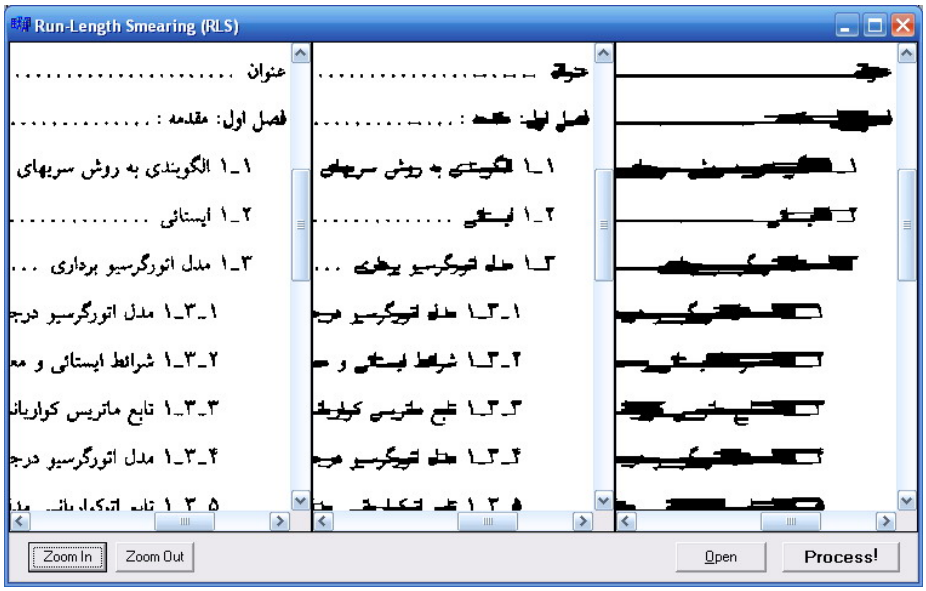
هدف از این مرحله، حذف اجزای زائد صفحه شامل قاب، خطوط، تصاویر و نویز صفحه است. عملیات پاکسازی در دو مرحله صورت می‌گیرد:

مرحلهٔ اول: هدف از این مرحله، حذف اجزای[۶۳] زائد بزرگ نظیر قاب، خطوط راست و منحنی، تصاویر و نویز غالب، و همچنین حذف اجزای ریز پراکنده در صفحه می‌باشد. برای این منظور ابتدا شیوهٔ «آغشته‌سازی»[۶۴] یا «آرال‌اس» در چهار جهت افقی، عمودی، ۴۵ درجه و ۱۳۵ درجه، بر روی تصویر اعمال می‌شود.

شیوهٔ آغشته‌سازی نخستین بار توسط «وانگ، کیسی و وال» مطرح گردید (Wong, Casey, & Wahl, 1982) . در این شیوهٔ صفحه در جهت مورد نظر (افقی، عمودی یا هر جهت دیگر) اسکن می‌شود و فضاهای سفید بین پیکسل‌های سیاه که طولشان کمتر از مقدار یک سطح آستانه[۶۵] است، سیاه می‌گردند. با این عمل، اجزای مجزای نزدیک به هم در تصویر به هم پیوند می‌خورند و اجزای متصل[۶۶] بزرگ‌تری را تشکیل می‌دهند. البته این شیوهٔ در صورتی مفید است که کجی تصویر اصلاح شده باشد و فاصله‌بندی‌ها (فواصل بین حروف کلمات، بین خود کلمات، بین سطور متن در یک پاراگراف، بین دو پاراگراف متوالی و بین دو ستون متن مجاور) مشخص و در کل تصویر یکنواخت باشد.

از آنجا که فاصلهٔ متوسط بین حروف در یک کلمه کم‌تر از فاصلهٔ متوسط بین دو کلمهٔ متوالی در متن است، سطح آستانهٔ افقی (برای آغشته‌سازی متن در جهت افقی) می‌تواند به نحوی انتخاب گردد که تنها حروف تشکیل‌دهندهٔ کلمات را به هم پیوند دهد، اما مانع از اتصال خود کلمات به یکدیگر شود (شکل ۶، تصویر وسط). از طرف دیگر با انتخاب بزرگ‌تر سطح آستانهٔ افقی می‌توان کلیهٔ کلمات موجود در یک خط از متن را به هم متصل نمود (شکل ۶، تصویر سمت راست). همچنین سطح آستانهٔ عمودی (برای آغشته‌سازی متن در جهت عمودی) را می‌توان کوچک انتخاب کرد، به نحوی که تنها سبب اتصال اجزای حروف به یکدیگر شود (مثلاً اتصال نقطه‌های حروف،

سرکش، علامت مد، تشدید و غیره به بدنهٔ حروف)، اما سطور متوالی متن را به یکدیگر پیوند ندهد. یا اینکه با انتخاب سطح آستانهٔ عمودی بزرگ‌تر می‌توان سبب اتصال کلیهٔ سطور متن به یکدیگر یا فقط باعث اتصال کلیهٔ سطور یک پاراگراف به یکدیگر گردید (به شرطی که فاصلهٔ بین سطور پاراگراف، کم‌تر از فاصلهٔ بین دو پاراگراف متوالی باشد).



شکل ۶ تصویر سمت چپ: یک صفحهٔ فهرست مطالب. تصویر وسط: آغشته‌سازی افقی در سطح حروف. تصویر سمت راست: آغشته‌سازی افقی در سطح کلمات

در این مرحله از عملیات پاکسازی، سطح آستانهٔ افقی به نحوی انتخاب می‌گردد که مانع از اتصال کلمات متن به یکدیگر و نیز قاب صفحه به متن شود. همچنین سطح آستانهٔ عمودی نیز به اندازه‌ای انتخاب شده است که اجزای حروف را به بدنهٔ حروف می‌چسباند، اما سطور متوالی متن را به هم اتصال نمی‌دهد.

پس از آغشته‌سازی تصویر، نوبت به پیدا کردن اجزای متصل در تصویر یا اصطلاحاً قطعه‌بندی تصویر می‌رسد. دو رویکرد زیر برای قطعه‌بندی تصویرقابل تصور است:

الف. تصویر را مورد پیمایش قرار می‌دهیم و به هر پیکسل سیاهی که برخورد کردیم، یا آن را به یکی از مجموعه اجزای متصل که قبلاً یافته شده‌اند اضافه می‌کنیم، یا برای

آن مجموعهٔ جدیدی در نظر می‌گیریم. در این روش تنها پس از پیمایش کل تصویر است که از کامل‌شدن هر یک از مجموعه‌ها اطمینان حاصل می‌شود.

ب. در پیمایش تصویر به محض پیدا شدن یک پیکسل سیاه جدید، جستجو را برای یافتن کلیهٔ پیکسل‌های سیاهی که با آن یک جزء متصل تشکیل می‌دهند ادامه می‌دهیم و وقتی این مجموعه کامل شد، به پیمایش تصویر برای یافتن پیکسل سیاه جدید بعدی می‌پردازیم. در این روش پس از یافتن اولین پیکسل سیاه متعلق به یک جزء متصل، ابتدا مجموعه پیکسل‌های آن را کامل می‌کنیم و بعد به دنبال کامل کردن دیگر مجموعه‌ها می‌رویم.

از آنجاکه ما در مرحلهٔ پاکسازی می‌خواهیم پس از یافتن هر جزء متصل، در مورد حذف یا عدم حذف آن تصمیم‌گیری نماییم، روش دوم را به کار برده‌ایم. ما یک الگوریتم جدید بازگشتی را برای پیدا کردن اجزای متصل در تصویر، معرفی و پیاده‌سازی نموده‌ایم که در بین فنون معرفی‌شده برای قطعه‌بندی تصاویر، در خانوادهٔ شیوهٔ «رشد ناحیه[۶۷]» (Gonzalez & Woods, 2002, pp. 613-615) قرار می‌گیرد. الگوریتم پیشنهادی ما به ترتیب زیر عمل می‌نماید:

۱. تصویر را در راستای افقی مورد پیمایش قرار می‌دهیم و به اولین پیکسل سیاهی که رسیدیم، آن را به عنوان نقطهٔ شروع برای یافتن کلیهٔ پیکسل‌هایی که به همراه آن، یک جزء پیوسته را تشکیل می‌دهند، در نظر می‌گیریم. به این پیکسل در اصطلاح پردازش تصویر، «بذر»[۶۸] گفته می‌شود. برای نگهداری پیکسل‌های جزء متصل به پیکسل بذر، مجموعه‌ای را تعریف می‌کنیم و پیکسل بذر را به عنوان اولین عضو به آن اضافه می‌نماییم. پیکسل بذر را به عنوان پیکسل «مرجع» برای پردازش‌های بعدی در نظر می‌گیریم.

۲. اگر در همسایگی مرتبهٔ ۸ پیکسل مرجع، پیکسل‌های سیاه جدیدی که قبلاً عضو مجموعه نبوده‌اند موجود باشند، آن‌ها را به مجموعه اضافه می‌کنیم. در غیر این صورت به مرحلهٔ ۶ می‌رویم.

۳. «مناسب‌ترین همسایهٔ» پیکسل مرجع را پیدا می‌کنیم و آن، پیکسلی است که در میان پیکسل‌های سیاه همسایهٔ پیکسل مرجع، در همسایگی خود دارای بیش‌ترین تعداد پیکسل‌های سیاه جدیدی باشد که قبلاً عضو مجموعه نبوده‌اند. دلیل آن‌که از بین

پیکسل‌های همسایهٔ پیکسل مرجع، مناسب‌ترین همسایه را برای ادامهٔ کار در نظر می‌گیریم آن است که با این پیکسل، احتمال آن‌که (نه با قطعیت) تعداد بازگشت‌های الگوریتم بازگشتی کم‌تر شود، بیش‌تر است.

۴. اگر مناسب‌ترین همسایه وجود داشته باشد، آن را به عنوان پیکسل مرجع در نظر می‌گیریم و به مرحلهٔ ۲ می‌رویم. در غیر این صورت بررسی می‌کنیم که آیا قبل از پیکسل مرجع کنونی، پیکسل مرجعی داشته‌ایم یا خیر.

۵. اگر پاسخ مرحلهٔ ۴ مثبت بود، به پیکسل مرجع قبلی برمی‌گردیم و مجدداً آن را به عنوان مرجع برمی‌گزینیم و سپس به مرحلهٔ ۳ می‌رویم. اگر پاسخ منفی بود، به مرحلهٔ ۶ رجوع می‌نماییم.

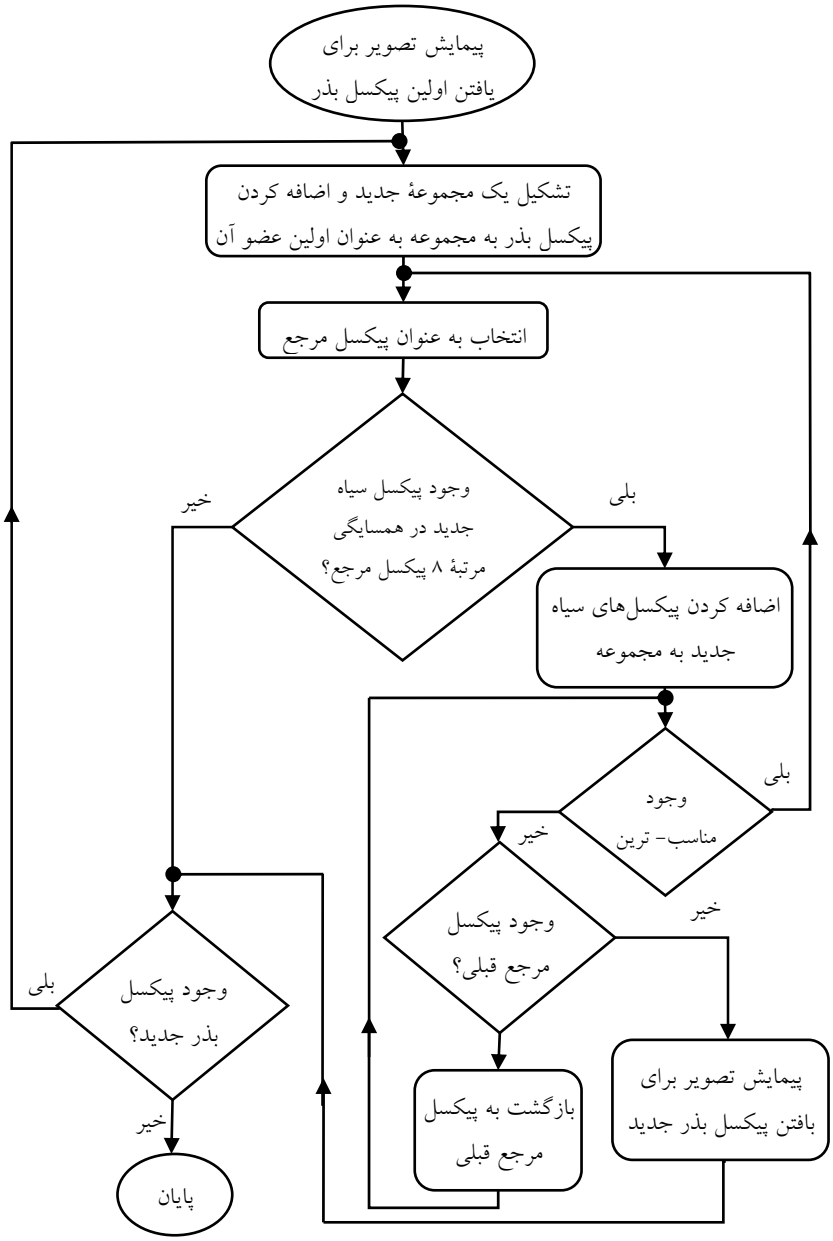
۶. از محل پیکسل بذر، تصویر را برای یافتن پیکسل بذر جدید پیمایش می‌کنیم و در صورت وجود، آن را به عنوان مرجع انتخاب می‌کنیم و به مرحلهٔ ۲ می‌رویم؛ در غیر این صورت، الگوریتم خاتمه می‌یابد. باید توجه داشت که در پیمایش برای یافتن پیکسل بذر جدید، پیکسل‌هایی که با یکی از پیکسل‌های بذر قبلی تشکیل یک جزء متصل می‌دادند را در نظر نمی‌گیریم.

شکل ۷ بلوک دیاگرام این روش را نشان می‌دهد.

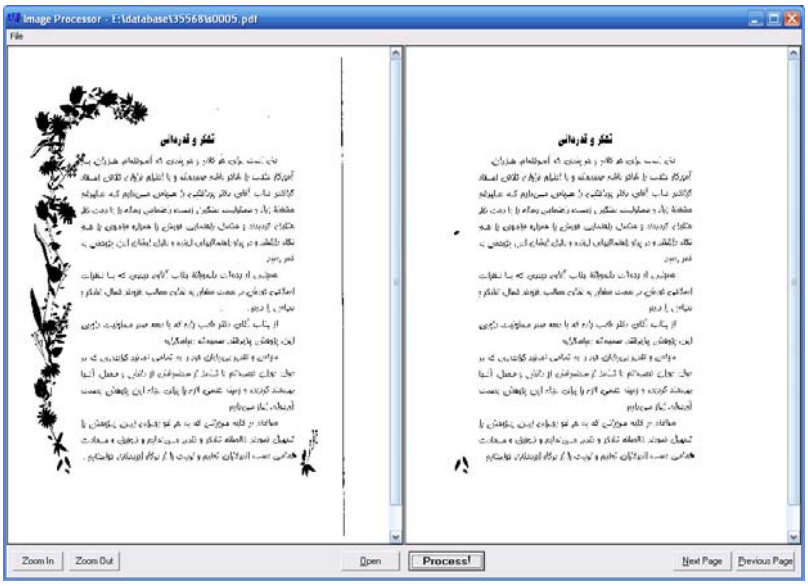
پس از اتمام عمل قطعه‌بندی، با انتخاب سطوح آستانهٔ مناسب، اجزایی را که اندازهٔ آن‌ها در جهت افقی یا عمودی بسیار بزرگ یا بسیار کوچک باشند حذف می‌نماییم.

مرحلهٔ دوم: در این مرحله از عملیات پاکسازی سعی می‌شود تا حد امکان، اجزای بسیار کوچک تصویر (اجزای نویزی ریز) که در مرحلهٔ اول حذف نشده‌اند، پاکسازی گردند. پس در این مرحله (به منظور اتصال نقاط حروف به بدنهٔ آن‌ها و ممانعت از حذف شدنشان)، شیوهٔ آغشته‌سازی تنها در جهت عمودی برروی تصویر اعمال می‌شود.

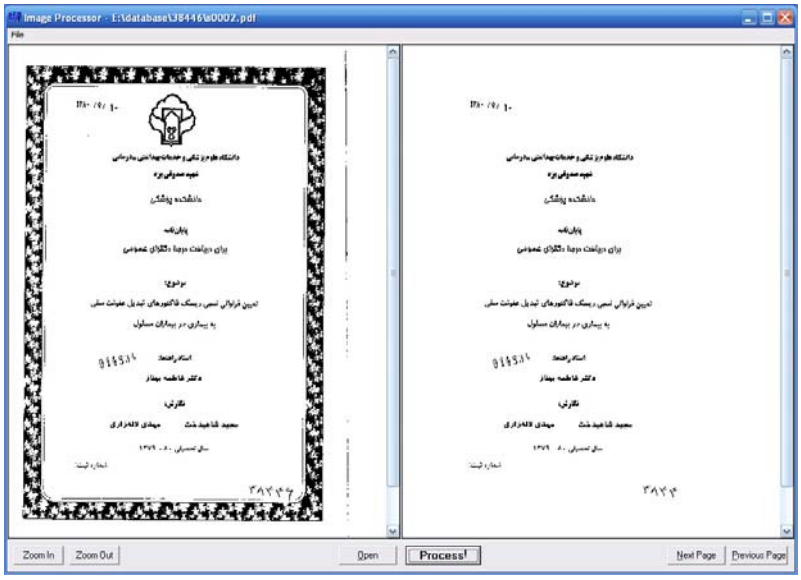
شکل‌های ۸ تا ۱۱ چند صفحهٔ نمونه از پایان‌نامه‌ها را قبل و بعد از اعمال عملیات پاکسازی به نمایش می‌گذارند. تصویر سمت چپ در هر شکل، تصویر اصلی را نشان می‌دهد و تصویر سمت راست، تصویر پاکسازی‌شدهٔ آن است. مأخذ پایان‌نامه‌ها، پایگاه اطلاعاتی پژوهشگاه اطلاعات و مدارک علمی ایران می‌باشد.



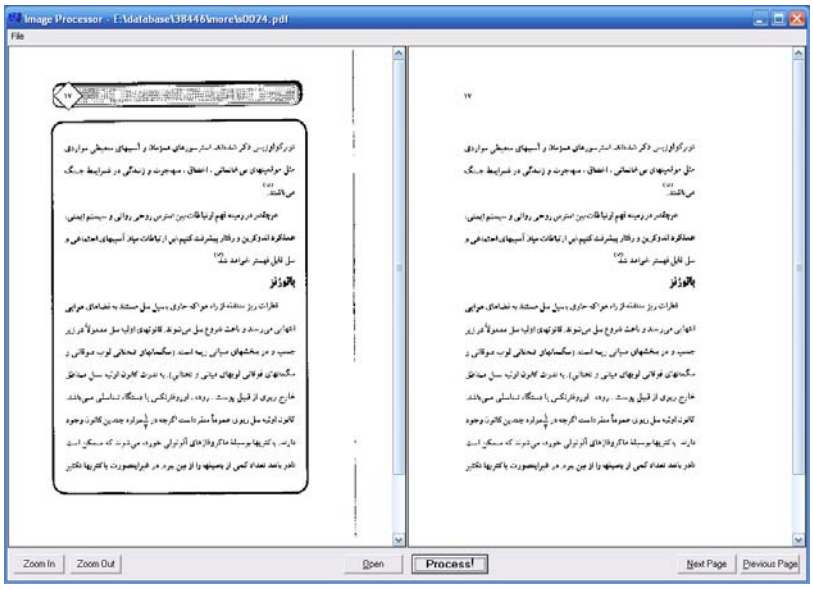
شکل ۷ بلوک دیاگرام الگوریتم پیشنهادی برای قطعه‌بندی تصویر



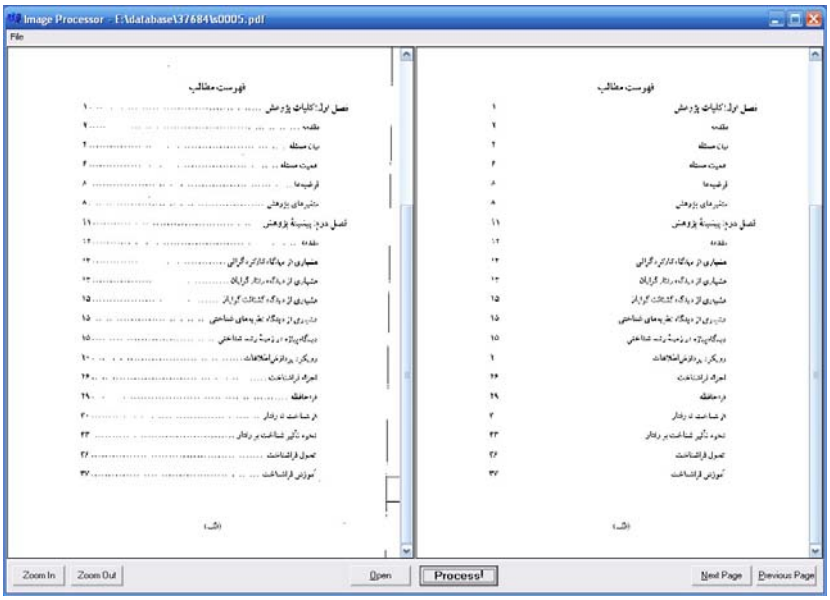
شکل ۸ صفحهٔ ۵ از پایان‌نامه به شمارهٔ کاربرگه (بازیابی) TH35568 (سمت چپ) و تصویر پاکسازی شدهٔ آن (سمت راست). اجزای زائد صفحه به خوبی حذف گردیده‌اند.



شکل ۹ صفحهٔ ۲ از پایان‌نامه به شمارهٔ کاربرگه (بازیابی) TH38446 (سمت چپ) و تصویر پاکسازی‌شدهٔ آن (سمت راست). اجزای زائد صفحه به خوبی حذف گردیده‌اند.



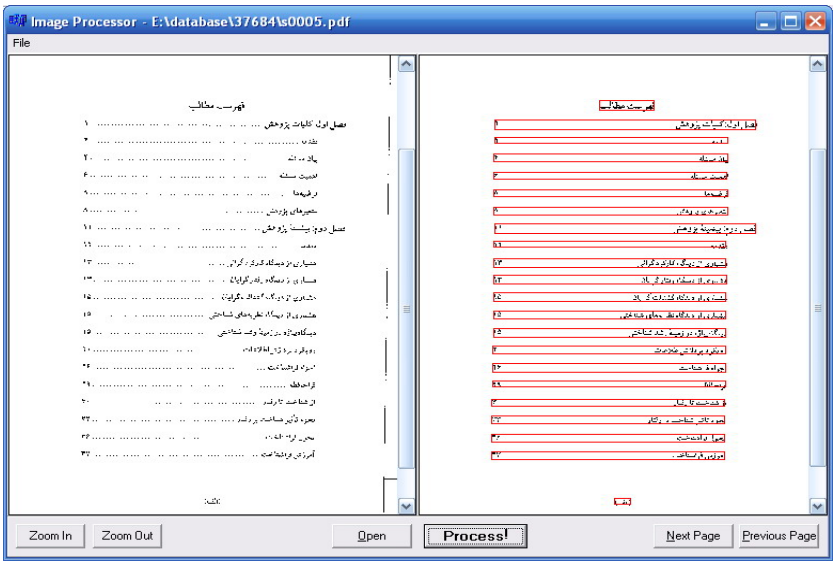
شکل ۱۰ صفحهٔ ۲۴ از پایان‌نامه به شمارهٔ کاربرگه (بازیابی) TH38446 (سمت چپ) و تصویر پاکسازی‌شدهٔ آن (سمت راست). اجزای زائد صفحه به خوبی حذف گردیده‌اند.



شکل ۱۱ صفحهٔ ۵ (فهرست مطالب) از پایان‌نامه به شمارهٔ کاربرگه (بازیابی) TH37684 (سمت چپ) و تصویر پاکسازی‌شدهٔ آن (سمت راست). سطور راهنما حذف گردیده‌اند.

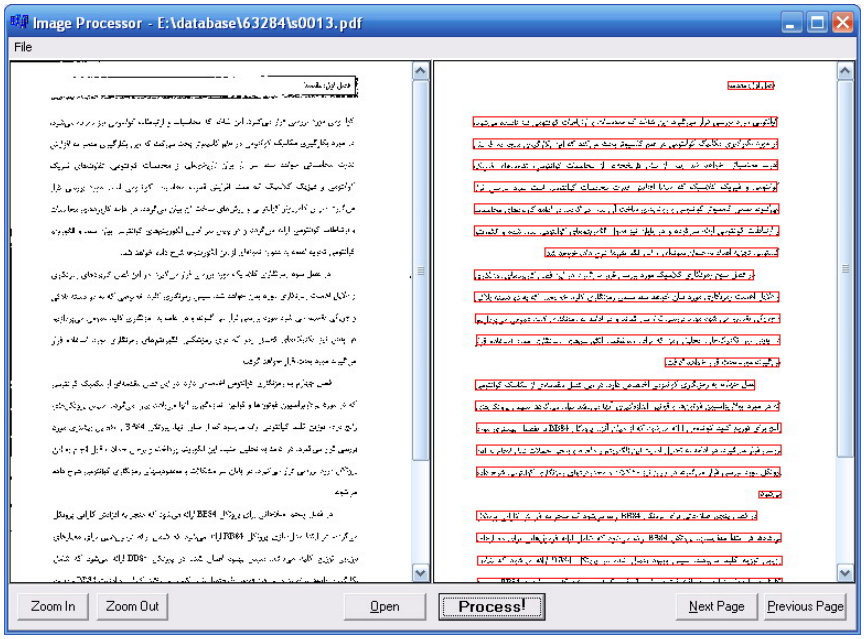
۴-۲. عملیات فریم‌بندی[۶۹]

هدف از این عملیات، یافتن فریم سطور متن در تصویر است. برای این منظور، ابتدا شیوهٔ آرال‌اس را در جهت‌های افقی و عمودی بر روی تصویر اعمال می‌کنیم. مقدار سطح آستانهٔ افقی باید قدری بزرگ‌تر از فاصلهٔ بین کلمات مجاور در هر سطر متن انتخاب گردد. سپس با محاسبهٔ هیستوگرام افقی صفحه، حدود بالایی و پایینی فریم‌های متن مشخص می‌شوند. به‌منظور افزایش دقت، میله‌هایی از هیستوگرام[۷۰] که مقادیرشان زیر یک سطح آستانهٔ کوچک باشند، حذف می‌شوند تا اثر نویزهای احتمالی باقیمانده کاهش یابد. همچنین چنانچه فاصلهٔ بین حدود بالایی و پایینی هر فریم یافته‌شده کم‌تر از حداقل ارتفاع مورد انتظار برای یک سطر از متن باشد، آن فریم در نظر گرفته نمی‌شود. فعلاً حدود چپ و راست فریم‌های متنی را به ترتیب، لبهٔ چپ و راست تصویر در نظر می‌گیریم.



شکل ۱۲ صفحهٔ ۵ (فهرست مطالب) از پایان‌نامه به شمارهٔ کاربرگه (بازیابی) TH37684
(سمت چپ) و تصویر فریم‌بندی شدهٔ آن (سمت راست)

برای محاسبهٔ دقیق حدود چپ و راست فریم‌های سطور متن، هیستوگرام عمودی هر فریم را به‌دست می‌آوریم.حد سمت چپ فریم، مکان سمت‌چپ‌ترین میلهٔ هیستوگرام عمودی با مقدار غیر صفر است به‌شرطی که حداقل ۵ میلهٔ مجاور سمت راست آن، مقادیر غیر صفر داشته باشند. به همین ترتیب، حد سمت راست فریم، مکان سمت‌راست‌ترین میلهٔ هیستوگرام عمودی با مقدار غیر صفر است به‌شرطی که حداقل ۵ میلهٔ مجاور سمت چپ آن، مقادیر غیر صفر داشته باشند. شکل‌های ۱۲ و ۱۳ عملیات فریم‌بندی را برای چند صفحهٔ نمونه از پایان‌نامه‌ها به نمایش می‌گذارند.



شکل ۱۳ صفحهٔ ۱۳ از پایان‌نامه به شمارهٔ کاربرگه (بازیابی) TH63284 (سمت چپ) و تصویر فریم‌بندی شدهٔ آن (سمت راست)

۴-۳. تعیین سطرهای کاندیدای فهرست

در صفحهٔ فهرست، سطرهایی از متن را که شامل عنوان و شمارهٔ صفحهٔ مربوط به یکی از بخش‌های پایان‌نامه باشند سطور فهرست می‌نامیم. در این مرحله ابتدا به جستجوی سطور کاندیدای فهرست در صفحه می‌پردازیم. تصویر ورودی همان تصویر آرال‌اس شدهٔ

مرحلهٔ قبل است. برای یافتن سطور کاندیدای فهرست، ابتدا طول کلیهٔ فریم‌های یافته‌شده در مرحلهٔ قبل را محاسبه، و آن‌ها را به‌ترتیب نزولی مرتب می‌کنیم. سپس بزرگ‌ترین فریم را در نظر می‌گیریم و تعداد فریم‌هایی را که اختلاف طولشان با طول فریم مبنا از یک حد آستانه کم‌تر باشند شمارش می‌کنیم. چنانچه تعداد این فریم‌ها بیش از نصف تعداد کل فریم‌های صفحه باشد، این طول فریم را به عنوان طول ماکزیمم سطرهای متن در نظر می‌گیریم و آن را طول مبنا می‌نامیم. در غیر این صورت، عملیات شمارش فوق را مجدداً برای بزرگ‌ترین فریم بعدی تکرار می‌کنیم. این کار را آنقدر ادامه می‌دهیم تا طول مبنا تعیین شود. چنانچه طول مبنا یافت نشود (یعنی هیچ طول فریمی پیدا نشود که مقدار فراوانی آن، حداقل نصف تعداد کل فریم‌ها باشد)، یا طول مبنا کم‌تر از ۴۰ درصد عرض صفحه باشد، این صفحه کاندید صفحهٔ فهرست نخواهد بود.

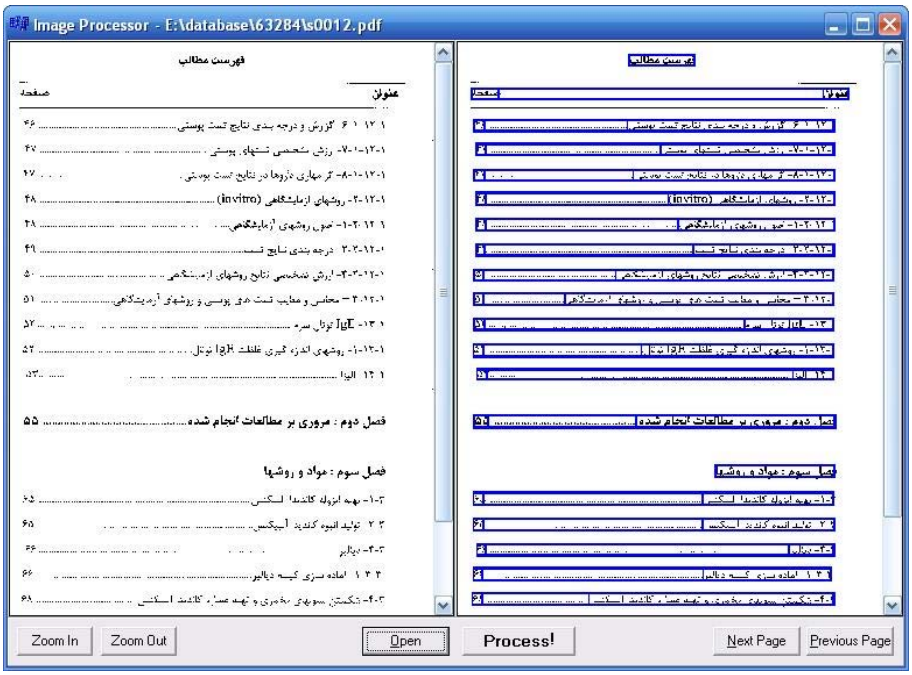
تا اینجا، فریم‌هایی از صفحه را که اختلاف طولشان با طول مبنا کم‌تر از یک سطح آستانه است به عنوان کاندیداهای اولیه برای سطور فهرست مشخص کرده‌ایم. حال از بین این فریم‌ها، فریم‌هایی را که کاندیدای نهایی هستند تعیین می‌نماییم. روش کار به صورت زیر است:

تفاوت بین یک سطر متن معمولی با یک سطر فهرست در این است که در یک سطر متن معمولی، پس از آرال‌اس شدن با یک سطح آستانهٔ کمی بزرگتر از حداکثر فاصلهٔ معمول بین کلمات متن، کلیهٔ کلمات به یکدیگر می‌چسبند. اما در یک سطر فهرست به دلیل فاصلهٔ زیاد بین عنوان و شمارهٔ صفحه، این دو بخش به هم متصل نمی‌شوند و در نتیجه فریم سطر فهرست، متشکل از دو زیرفریم[۷۱] است. البته چنانچه سطرهای راهنما که عناوین متن را به شمارهٔ صفحات ارتباط می‌دهند در صفحهٔ فهرست وجود داشته و در مرحلهٔ پاکسازی نیز حذف نشده باشند، دو زیرفریم به هم اتصال پیدا می‌کنند. به همین دلیل باید این سطور راهنما شناسایی و حذف گردند. برای این منظور، ما هیستوگرام عمودی هر فریم را محاسبه می‌کنیم و در صورتی‌که قطعه‌ای از هیستوگرام با حداقل طول α یافت بشود که برای کلیهٔ نقاط آن قطعه، مقدار ارتفاع هیستوگرام کم‌تر از β باشد، این قطعه مربوط به تصویر آرال‌اس شدهٔ خطوط راهنما است و بنابراین،

ما تصویر متناظر با این قسمت را از فریم حذف می‌نماییم. پس از حذف خطوط راهنما، زیر فریم‌ها به راحتی در فریم تشخیص داده می‌شوند.

شکل ۱۴ یک صفحهٔ فهرست را نشان می‌دهد که در آن، زیرفریم‌ها مشخص گردیده‌اند.

به عنوان آخرین آزمون، اگر طول یکی از زیرفریم‌ها (زیرفریم شمارهٔ صفحه - این زیرفریم در صفحهٔ فهرست فارسی یا عربی، زیرفریم سمت چپ است و در صفحهٔ فهرست لاتین، زیرفریم سمت راست می‌باشد) کم‌تر از γ و فاصلهٔ بین دو زیرفریم، کم‌تر از δ باشد، آن فریم کاندید نهایی برای یک سطر فهرست خواهد بود. مقدار γ باید آنقدر بزرگ باشد که یک سطر فهرست حاوی شماره صفحهٔ محدوده‌ای (به بخش ۳ مراجعه شود) نیز قابل تشخیص باشد. توجه شود که این روش، به ترازبندی یا جهت متن صفحه وابسته نیست.



شکل ۱۴ صفحهٔ ۱۱ از پایان‌نامه به شمارهٔ کاربرگه (بازیابی) TH39993 (سمت چپ) و تصویر فریم‌بندی‌شدهٔ آن (سمت راست). زیرفریم‌ها نیز مشخص گردیده‌اند.

۴-۴. تأیید نهایی

هر صفحه‌ای از پایان‌نامه که رابطهٔ زیر در مورد آن صادق باشد، به عنوان صفحهٔ فهرست تشخیص داده می‌شود:

$$\frac{\text{تعداد خطوط کاندیدای فهرست}}{\text{تعداد کل خطوط صفحه}} > 0.4$$

یعنی چنانچه تعداد سطرهای کاندیدای فهرست از ۴۰ درصد تعداد کل سطرهای صفحه بیش‌تر باشد، آن صفحه یک صفحهٔ فهرست می‌باشد. عدد ۴۰ درصد به این دلیل انتخاب شده است که در یک صفحهٔ فهرست، حتی اگر نام کلیهٔ عناوین به حدی طولانی باشند که شمارهٔ صفحهٔ مربوط به آن‌ها در سطور جداگانه‌ای قرار بگیرند (یعنی نصف خطوط صفحه شامل شمارهٔ صفحه نباشند)، باز هم آن صفحه به عنوان صفحهٔ فهرست تشخیص داده شود.

۵. آزمایش روش پیشنهادی

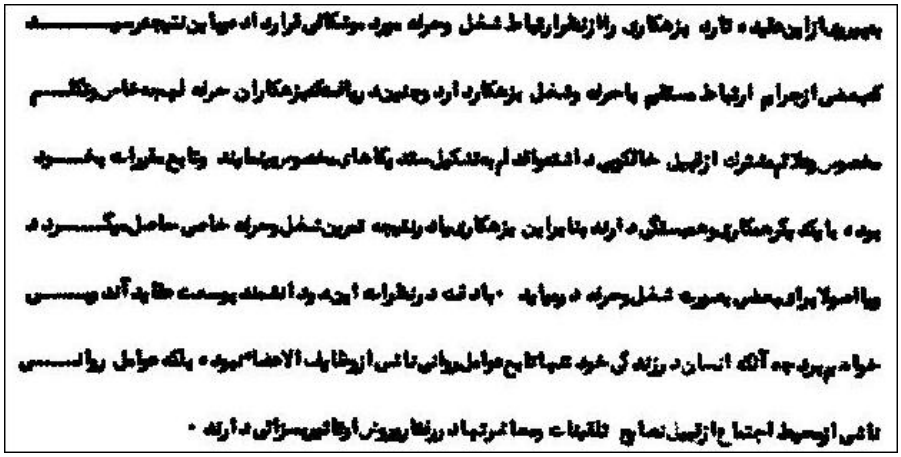
پایگاه اطلاعاتی پژوهشگاه اطلاعات و مدارک علمی ایران شامل پایان‌نامه‌های کارشناسی ارشد و دکترای دفاع‌شده در دانشگاه‌های داخل کشور از سال ۱۳۱۰ به بعد می‌باشد. در حال حاضر حدود ۹۰ هزار پایان‌نامه به طور کامل به فرم الکترونیکی تبدیل شده‌اند که اطلاعات کتاب‌شناسی به همراه ۱۵ صفحهٔ اول آن‌ها از طریق وب‌سایت پژوهشگاه برای عموم قابل دسترس می‌باشد.

به منظور آزمایش روش پیشنهادی، ۲۰۰ پایان‌نامه (مجموعاً دارای بیش از ۳۰ هزار صفحه) را به صورت تصادفی از پایگاه اطلاعاتی برگزیدیم. انتخاب به نحوی انجام شد که قدیمی‌ترین و جدیدترین پایان‌نامه‌ها را دربرگیرد. کیفیت نسخهٔ کاغذی برخی از این پایان‌نامه‌ها (که نسخهٔ الکترونیکی از اسکن کردن آن‌ها به دست آمده است) در گذر زمان آنچنان تنزل پیدا کرده است که حتی توسط انسان نیز به زحمت قابل خواندن می‌باشند (شکل ۱۵)، چه رسد به این‌که یک سیستم اُسی‌آر بخواهد متن آن‌ها را استخراج نماید. اما همان‌گونه که قبلاً عنوان گردید، به جهت آن‌که اطلاعات ساختاری (نظیر محل فریم خطوط متن، محل زیرفریم‌های عنوان و شمارهٔ صفحات، ...) تقریباً از

کیفیت مدارک تأثیر نمی‌پذیرد، روش معرفی‌شده در این مقاله قادر است به صورت صحیح صفحات فهرست را تشخیص دهد.

با آزمایش مشخص گردید که روش پیشنهادی در این مقاله قادر است ۹۹/۷ درصد از صفحات را به درستی به عنوان صفحهٔ فهرست تشخیص دهد. موارد خطا تنها مربوط به برخی پایان‌نامه‌های بسیار قدیمی بود که صفحات فهرست آن‌ها از مدل عنوان‌شده در بخش ۳ تبعیت نمی‌کردند. همچنین ۰/۰۶ درصد از صفحات به غلط به عنوان صفحهٔ فهرست تشخیص داده شدند که بیش‌تر آن‌ها، صفحاتی شامل یک جدول با ساختاری دقیقاً مشابه ساختار یک صفحهٔ فهرست بودند. با توجه به این‌که صفحات فهرست در پایان‌نامه‌ها در ۲۰ صفحهٔ اول قرار دارند، با محدود کردن جستجو تنها در این محدوده می‌توان سرعت و دقت الگوریتم را باز هم بهبود بخشید.

نتایج حاصل از آزمایش روش پیشنهادی، دقت بسیار بالای آن را در شناسایی صفحات فهرست در پایان‌نامه‌ها مورد تأیید قرار می‌دهد.



شکل ۱۵ نمونه‌ای از متن یک پایان‌نامه دارای کیفیت بسیار نازل

۶. جمع‌بندی و نتیجه‌گیری

در این مقاله برای اولین بار روشی به منظور شناسایی خودکار صفحات فهرست در پایان‌نامه‌های فارسی، عربی و لاتین ارائه شده است. در این روش تنها از اطلاعات

ساختاری صفحات فهرست که توسط نویسندگان پایان‌نامه‌ها لحاظ می‌شوند استفاده گردیده است و بر خلاف اکثر روش‌های پیشنهادی، اطلاعات متن صفحات مورد استفاده قرار نمی‌گیرد. از آنجا که عملیات استخراج متن از صفحات اسکن‌شده (اُسی‌آر)، بار محاسباتی نسبتاً زیادی دارد، این روش نسبت به کلیهٔ روش‌های مبتنی بر اُسی‌آر از سرعت بیش‌تری برخوردار است و بر خلاف آن‌ها، در صفحات دارای کیفیت چاپ یا اسکن بسیار پایین هم قابل به‌کارگیری می‌باشد. اساساً به جهت آن‌که در حال حاضر یک موتور اُسی‌آر فارسی با کارایی و قیمت مناسب در بازار موجود نیست، روش‌های شناسایی صفحات فهرست که مبتنی بر متن هستند، بر روی مدارک فارسی به راحتی قابل اعمال نیستند.

دقت بسیار بالای این روش، کارآیی آن را در شناسایی صفحات فهرست پایان‌نامه‌ها مورد تأیید قرار می‌دهد. هر چند روش مذکور تنها بر روی پایان‌نامه‌ها آزمایش گردیده است، اما برای مدارکی که صفحات فهرست در آن‌ها ساختاری مشابه با پایان‌نامه‌ها دارند (کتاب‌ها، مجلات، ...) نیز می‌توان از آن استفاده کرد.

روش معرفی‌شده در این مقاله در نرم‌افزاری با عنوان «پردازشگر مدارک» یا IDP[۷۲] مورد استفاده قرار گرفته است (فرامرزی، ۱۳۸۴، ۱۳۸۵). این نرم‌افزار پس از شناسایی صفحات فهرست، توسط یک الگوریتم پیشنهادی اُسی‌آر اعداد فارسی و لاتین، شمارهٔ صفحات عناوین بخش‌ها در صفحات فهرست را تشخیص می‌دهد؛ سپس توسط یک الگوریتم دیگر، صفحهٔ اول هر پایان‌نامه مشخص می‌شود. به دلیل وجود صفحات بدون شماره یا با شماره‌گذاری متفاوت در ابتدای هر پایان‌نامه، صفحهٔ دارای شمارهٔ یک، اولین صفحهٔ پایان‌نامه نخواهد بود. در نهایت، این نرم‌افزار بین صفحات فهرست و صفحات اصلی پایان‌نامه پیوند برقرار می‌کند. در نتیجه کاربر با کلیک بر روی هر یک از عناوین صفحات فهرست خواهد توانست مستقیماً به صفحه مربوطه در پایان‌نامه مراجعه نماید.

۷. منابع

۱. پایگاه اطلاعاتی ایران‌داک: پایگاه اطلاعاتی پژوهشگاه اطلاعات و مدارک علمی ایران. دسترسی در ۲۶ آبان ۱۳۸۵. از سایت http://database.irandoc.ac.ir/

۲. عزمی، رضا. (۱۳۷۸). بازشناسی متون چاپی فارسی. پایان‌نامهٔ دکترای مهندسی برق و الکترونیک، دانشگاه تربیت مدرس، *دانشکدهٔ فنی و مهندسی*.

۳. فرامرزی، اسماعیل. (۱۳۸۵). *ارائهٔ نسخهٔ دوم نرم‌افزار پردازشگر مدارک (طرح پژوهشی)*. تهران: پژوهشگاه اطلاعات و مدارک علمی ایران.

۴. فرامرزی، اسماعیل. (۱۳۸۴الف). ارائهٔ نرم افزاری بمنظور استخراج متن چکیدهٔ پایان‌نامه‌های اسکن شده و نیز ایجاد ارتباط بین عناوین بخش فهرست مطالب و صفحات اصلی متناظر آن‌ها از طریق شناسایی خودکار یا نیمه خودکار شمارهٔ صفحات (طرح پژوهشی). تهران: پژوهشگاه اطلاعات و مدارک علمی ایران.

۵. فرامرزی، اسماعیل. (۱۳۸۴ب). بازشناسی نوری حروف: مروری بر مباحث نظری و ملاحظات کاربردی با تاکید بر مسائل خاص زبان فارسی. *فصلنامه علوم اطلاع‌رسانی*، ۲۰(۳و۴)، ۶۱-۳۳.

۶. فرامرزی، اسماعیل. (۱۳۸۳). *فاز مطالعاتی و امکان‌سنجی شناسایی شمارهٔ صفحات پایان‌نامه‌ها در بخش فهرست مطالب و صفحات اصلی جهت زمینه‌سازی برقراری ارتباط خودکار بین آن‌ها (طرح پژوهشی)*. تهران: *پژوهشگاه اطلاعات و مدارک علمی ایران*.

7. Arıca, N., & Yarman-Vural, F. T. (2001). An overview of character recognition focused on off-line handwriting. *IEEE Transactions on Systems, Man, and Cybernetics - Part C: Applic and Rev*, 31(2), 216-233.

8. Belaid, A., Pierron, L., & Valverde, N. (2000). Part-of-speech tagging for table of contents recognition. In *Proceedings. ICPR 15th International Conference: Vol. 4. Pattern recognition* (pp. 4451–4454). Barcelona, Espagne: IEEE computer society press.

9. Gonzalez, R. C. & Woods, R. E. (2002). *Digital Image Processing (2nd ed)*. New Jersey: Prentice Hall.

10. He, F., Ding, X., & Peng L. (2004). Hierarchical logical structure extraction of book documents by analyzing tables of contents. In S. Barney, H. Elisa, J. Hu, J. Allan (Eds.), *Document Recognition and Retrieval XI: Vol. 5296. Proceedings of the SPIE* (pp. 6-13). SPIE-International society for optical engine.

11. *ITU-T: International Telecommunication Union, Information technology - Open Document Architecture (ODA) and interchange*

format. Retrieved November 17, 2006, from http://www.itu.int/rec/T-REC-t

12. Kasturi, R., O'Gorman, L., & Govindaraju, V. (2002). Document image analysis: a primer. *S¯adhan¯a,* 27(Part 1), 3–22.
13. Le Bourgeois, F., Emptoz, H., & Souafi Bensafi, S. (2001). Document understanding using probabilistic relaxation: application on tables of contents of periodicals. In *Proceedings of the Sixth International Conference on Document Analysis and Recognition (ICDAR'01)* (pp. 508-512). Washington, DC, USA: IEEE computer society.
14. Lin, X. (2003). Text-mining based journal splitting. In *Proceedings of 7th International Conference on Document Analysis and Recognition (ICDAR'03)* Vol. 2 (pp. 1075-1079). Washington, DC, USA: IEEE Computer Society.
15. Lin, C. C., Niwa, Y., & Narita, S. (1997). Logical structure analysis of book document image using contents information. In *Proceedings of the fourth international conference on Document analysis and recognition* Vol. 2 (pp. 1048-1051). Washington, DC, USA: IEEE Computer Society.
16. Lin, X., & Xiong, Y. (2006). Detection and analysis of table of contents based on content association. *International Journal on Document Analysis and Recognition*, 8(2-3), 132-143.
17. Mandal, S., Chowdhury, S. P., Das, A. K., & Chanda, B. (2003). Automated detection and segmentation of table of contents page and index pages from document images. In *Proceedings of 12th International Conference on Image Analysis and Processing* (pp. 213-218). Washington, DC, USA: IEEE Computer Society.
18. Mori, S., Suen, C. Y., & Yamamoto, K. (1992). Historical review of OCR research and development. *Proceedings of IEEE,* 80(7), 1029–1058.
19. Nagy, G. (2000). Twenty years of document image analysis in PAMI. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22 (1), 38-62.

OASIS: Open Document Format for Office Applications (OpenDocument) TC, Developing an XML-based file format specification for office applications. Retrieved November 17,

2006, from http://www.oasis-open.org/committees/tc_home.php?wg_abbrev=office

20. O'Gorman, L. (1995). *Document image analysis* (R. Kasturi, Ed.). Los Alamitos, CA, USA: IEEE Computer Society Press.
21. O'Gorman, L. (1992). Image and document processing techniques for the right pages electronic library system. *Proceeding of 11th International Conference on Pattern Recognition (IAPR)* Vol. II. (pp. 260-263). Los Alamitos, California, USA: IEEE Computer Society Press.
22. Satoh, S., Takasu, A., & Katsura, E. (1995). An automated generation of an electronic library based on document image understanding. *Proceedings of the Third International Conference on Document Analysis and Recognition (ICDAR)* Vol. 1. (pp. 163-166). Washington, DC, USA: IEEE Computer Society.
23. Schurmann, J., Bartneck, N., Bayer, T., Franke, J., Mandler, E., & Oberlander, M. (1992). Document analysis- from pixels to contents. *Proceedings of the IEEE*, 80 (7), 1101-1119.
24. Srihari, S., Stephen, L., Venu, G., Rohini S., & Jonathan, H. (1992). *Document understanding: Research directions (Tech. Rep. CEDAR-TR-92-1)*. USA: State University of New York at Buffalo, Center of Excellence for Document Analysis and Recognition (CEDAR).
25. Story, G. A., O'Gorman, L., Fox, D., Schaper, L. L., & Jagadish, H. V. (1992). The right pages image-based electronic library for alerting and browsing. *IEEE Computer*, 25 (9), 17–26.
26. Takasu, A., Satoh, S., & Atsura, E. (1994). A document understanding method for database construction of an electronic library. *Proceedings of the 12th international conference on pattern recognition (ICPR)* Vol. 2 (pp. 463-466). Los Alamitos, California, USA: IEEE Computer Society Press.
27. Takasu, A., Satoh, S., & Katsura, E. (1995). A rule learning method for academic document image processing. *Proceedings of the Third International Conference on Document Analysis and Recognition (ICDAR'95)* Vol. 1. (pp. 239-242). Washington, DC, USA: IEEE Computer Society.

28. Tappert, C. C., Suen C. Y., & Wakahara, T. (1990). The state of the art in online handwriting recognition. *IEEE Transaction on Pattern Analysis and Machine Intelligence (PAMI)*, 12(8), 787–808.

29. Tsuruoka, S., Hirano, C., Yoshikawa, T., & Shinogi, T. (2001). Image-based structure analysis for a table of contents and conversion to xml documents. *Proceedings of 2nd Document Layout Interpretation and its Applications (DLIA2001)* (pp.59-62).

30. *W3C: The World Wide Web Consortium, Overview of SGML resources.* Retrieved November 17, 2006, from http://www.w3.org/MarkUp/SGML/

31. *W3C: The World Wide Web Consortium, The Extensible Stylesheet Language family (XSL).* Retrieved November 17, 2006, from http://www.w3.org/Style/XSL/

32. Wieser, J., & Pinz, A. (1993). Layout and analysis: Finding text, titles, and photos in digital images of newspaper pages. *Proceedings of the 2nd International Conference Document Analysis and Recognition (ICDAR)* (pp. 774-777). Washington, DC, USA: IEEE Computer Society.

33. *Wikipedia: multilingual web-based free content encyclopedia.* Retrieved November 17, 2006, from http://en.wikipedia.org

34. Wong, K. Y., Casey, R. G., & Wahl, F. M. (1982). Document analysis system. *IBM Journal of Research and Development*, 26(6), 647-656.

35. Yacoub, S., & Peiro, J. A. (2005). Identification of document structure and table of content in magazine archives. In B. Werner (Ed.), *Proceedings of the Eighth International Conference on Document Analysis and Recognition: Vol. 2. (pp. 1253- 1257).* Washington, DC: IEEE Computer Society.

پی‌نوشت‌ها

1. alignment
2. Document Image Analysis (DIA)
3. page layout analysis
4. structural analysis

5. logical analysis
6. document image understanding
7. contents page recognition
8. image processing
9. Optical Character Recognition (OCR)
10. Pattern recognition

۱۱. برای اطلاعات بیش‌تر در خصوص مباحث فنی، به (فرامرزی ۱۳۸۳ و ۱۳۸۴) مراجعه شود.

12. queries
13. graphical processing or graphics Processingprocessing
14. line fitting
15. picture processing or photographic processing
16. compression

۱۷. برای اطلاعات بیش‌تر در این خصوص به (فرامرزی، ۱۳۸۳، ۱۳۸۴؛ عزمی، ۱۳۷۸، Tappert, Suen, & Wakahara, 1990; Mori, Suen, & Yamamoto, 1992; Arıca & Yarman-Vural, 2001) مراجعه شود

18. syntactical
19. functional
20. segmentation
21. top-Down
22. bottom-Up
23. domain Dependant
24. labeling
25. indication
26. splitting
27. merging

۲۸. برای کسب اطلاعات بیش‌تر در حوزهٔ دی‌آی‌ای به (فرامرزی، ۱۳۸۳)، Kasturi, O'Gorman, & Govindaraju, 2002; Nagy, 2000; O'Gorman, 1995, pp. 161-181; Schurmann, et al., 1992) مراجعه شود.

29. Open Document Architecture or Office Document Architecture (ODA)
30. Standard generalized Markup Language (SGML)
31 . HTML (Hyper-Text Markup Language)
32. Cascading Style Sheets (CSS)
33 . XML (Extensible Markup Language)
34. Extensible Style sheet Language (XSL)
35. Open Document or ODF, short for the "OASIS Open Document Format for Office Applications"

۳۶. برای اطلاعات بیش‌تر در زمینهٔ هر یک از این استانداردها، علاوه بر مراجع فوق به (Wikipedia, Website) مراجعه گردد.

37. Indexing
38. link
39. Bookmark
40. Meta-structure
41. symbols
42. math-zones
43. cursive script
44. Docstrum
45. a priori model
46. tagging
47. POS (Part Of Speech)
48. regularity
49. image-based
50. indentation
51. probabilistic relaxation
52. right-alignment
53. decision tree
54. text-mining
55. matching
56. Digital Content Re-Mastering (DCRM)
57 . MIT Press
58. content association
59. bottom-up clustering
60. tabular structure
61. formatting
62. image cleaning
63. components
64. Run-Length Smearing (RLS)
65. threshold
66. connected components
67. region growing
68. seed
69. framing
70. histogram bins
71. sub-frame
72. IRANDOC Document Processor

۱) عضو هیئت علمی پژوهشگاه اطلاعات و مدارک علمی ایران
پست‌الکترونیکی: faramarzi@irandoc.ac.ir