A Structure-Based Method for Building a Database of Extracted Figures from Scientific Documents: A Case Study of Iran Scientific Information Database (GANJ)

Azadeh Fakhrzadeh*
PhD in Digital Image Processing; Assistant Professor;
Iranian Research Institute for Information Science and Technology (IranDoc); Tehran, Iran Email: Fakhrzadeh@irandoc.ac.ir

Amir Hossein Seddighi
PhD in Industrial Engineering; Assistant Professor;
Iranian Research Institute for Information Science and Technology (IranDoc); Tehran, Iran Email: Seddighi@irandoc.ac.ir

Iranian Journal of Information Processing and Management

Received: 16, Jan. 2018 **Accepted: 13, May 2018**

Iranian Research Institute for Information Science and Technology (IranDoc)
ISSN 2251-8223
eISSN 2251-8231
Indexed by SCOPUS, ISC, & LISTA
Vol. 35 | No. 3 | pp. 729-754
Spring 2020

Abstract: Figures in scientific documents are rich sources of information. The first step in retrieving information from such figures is to build a valid figure database. To this end, we developed a system for generating figure database from scholarly Persian documents, in large scale. The first step is to parse files and extract figures and their corresponding descriptions. There are two general approaches for extracting figures from documents. One is based on image processing methods and another is based on processing the file primitives. The focus of this paper is on latter one. This approach is shown to be a better choice for the search engines because of its speed and scalability properties. We propose a structure based method that extracts the figures and their descriptions by analyzing the file layout. This information is saved in a database with a specific structure and is indexed for retrieval in the search engine.

The proposed algorithm was implemented in Python programming language. As a benchmark we used the basic method in the literature which is based on the processing PDF file. We employed the proposed method in a case study on Iran scientific information database (Ganj). In this regard, 150 scientific documents were randomly chosen from Ganj database and analyzed using two mentioned methods. Based on our experimental results, the proposed method is more efficient than the basic method especially for Persian documents. There are many unanswered challenges for Persian documents when using the basic

* Corresponding Author

method. The number of noise images resulted from the basic method is high and Persian text extracted is not well organized. Our proposed method overcomes some of these drawbacks and is recommended for generating figure database from scientific Persian documents. The proposed method is able to correctly extract about 40% of the images with their corresponding descriptions which is 10% better than the basic method.

Keywords: Image Processing, Image Extraction, Metadata Extraction, Information Technology

ارائۀ روشی ساختارمحور برای ایجاد پایگاه داده از تصاویر مستخرج از اسناد علمی؛ مورد مطالعه: پایگاه اطلاعات علمی ایران (گنج)

آزاده فخرزاده
دکتری پردازش تصویر کامپیوتری؛ استادیار؛ پژوهشگاه علوم و فناوری اطلاعات ایران (ایرانداک)؛ تهران، ایران؛
پدیدآور رابط Fakhrzadeh@irandoc.ac.ir

امیرحسین صدیقی
دکتری مهندسی صنایع؛ استادیار؛ پژوهشگاه علوم و فناوری اطلاعات ایران (ایرانداک)؛ تهران، ایران Seddighi@irandoc.ac.ir

پژوهش‌نامۀ پردازش و مدیریت اطلاعات

دریافت: ۱۳۹۸/۰۱/۲۱ پذیرش: ۱۳۹۸/۰۸/۰۵ مقاله برای اصلاح به مدت ۵۵ روز نزد پدیدآوران بوده است.

نشریۀ علمی | رتبۀ بین‌المللی
پژوهشگاه علوم و فناوری اطلاعات ایران (ایرانداک)
شاپا (چاپی) ۸۲۲۳-۲۲۵۱
شاپا (الکترونیکی) ۸۲۳۱-۲۲۵۱
نمایه در LISTA، ISC، و SCOPUS
jipm.irandoc.ac.ir
دورۀ ۳۵ | شمارۀ ۳ | ص‌ص ۷۲۹-۷۵۴
بهار ۱۳۹۹

چکیده: تصاویر موجود در مدارک علمی اغلب حاوی اطلاعات مهمی هستند. اولین قدم برای بازیابی اطلاعات از این تصاویر ایجاد یک پایگاه دادۀ معتبر از آن‌هاست. به این منظور در این مقاله سیستمی خودکار برای ایجاد پایگاه داده از تصاویر موجود در مدارک علمی فارسی در مقیاس بزرگ ارائه می‌شود. این سیستم پیشنهادی به‌دنبال مطالعات اسنادی طراحی شده و بخش‌های مختلفی دارد. در مرحلۀ اول باید تصاویر و توضیح متنی آن‌ها استخراج گردد. به‌طور کلی، دو رویکرد برای استخراج تصاویر و توضیح متنی آن‌ها از فایل وجود دارد. در رویکرد اول، فایل به تصویر تبدیل می‌شود و از تکنیک‌های پردازش تصویر برای استخراج اطلاعات گرافیکی استفاده می‌شود. رویکرد دوم، بر اساس پردازش ساختار و آرایش خود فایل است. از آنجا که روش دوم از لحاظ سرعت و قابلیت مقیاس‌پذیری برای استفاده در موتورهای جست‌وجو مناسب‌تر است، تمرکز این مقاله بر روی روش دوم است. بدین‌ترتیب، برای استخراج تصاویر و توضیح متنی آن‌ها از یک روش ساختارمحور استفاده می‌شود که مبتنی بر چیدمان و آرایش فایل وُرد سند است. در نتیجه، مجموعه‌ای از تصاویر به همراه توضیحات و اطلاعات مربوط به آن‌ها به‌دست می‌آید که باید در یک پایگاه دادۀ تصاویر با ساختاری مشخص ذخیره گردند. سپس، این اطلاعات برای بازیابی و استفاده‌های آتی در یک موتور جست‌وجو نمایه خواهند شد. روش پیشنهادی در زبان برنامه‌نویسی «پایتون» پیاده‌سازی شد و برای ارزیابی

کارایی آن از روش مرسوم پردازش فایل «پی‌دی‌اف» اسناد کمک گرفته شد. سپس، روش پیشنهادی در یک مطالعهٔ موردی در «پایگاه اطلاعات علمی ایران (گنج)» به‌کار گرفته شد. تعداد ۱۵۰ مدرک علمی به تصادف از «پایگاه گنج» انتخاب شده و با کمک این دو روش مورد تجزیه و تحلیل قرار گرفت. بنا به یافته‌های پژوهش دیده می‌شود که استخراج اطلاعات متنی از فایل «پی‌دی‌اف» در زبان فارسی با چالش‌های زیادی روبه‌روست و نمی‌تواند خروجی مناسبی در این زمینه حاصل کند. از طرف دیگر، میزان تصاویر نامطلوب تولیدشده از فایل «پی‌دی‌اف» بسیار زیاد است که از کاربست‌پذیری آن در شرایط واقعی می‌کاهد. از این رو، روش پیشنهادی به‌عنوان گزینه‌ای مناسب برای استخراج تصویر و توضیحات آن‌ها از اسناد علمی در زبان فارسی و ایجاد پایگاه داده از آن‌ها پیشنهاد می‌شود. روش پیشنهادی قادر است حدود ۴۰ درصد تصاویر را همراه با زیرنویس مربوطه بدون خطا استخراج کند و نسبت به روش پایه که قادر به استخراج ۳۰ درصد از تصاویر است، کارایی بهتری دارد.

کلیدواژه‌ها: پردازش تصویر، استخراج تصویر، استخراج فراداده، فناوری اطلاعات

۱. مقدمه

با فراهم‌آمدن امکان دیجیتال‌کردن اسناد، پیدا کردن روش‌های مؤثر و خودکار برای بازیابی اطلاعات از اسناد مورد توجه قرار گرفته است. بیشتر تلاش‌ها در این زمینه در راستای فهمیدن متن اسناد و استخراج اطلاعات فرادادهٔ متنی است. امروزه، با پیشرفت به‌وجودآمده در ایجاد تصاویر دیجیتال، تصاویر نیز به بخش مهمی از اسناد علمی تبدیل شده‌اند و می‌توانند خلاصهٔ مفیدی از این نوع اسناد را ارائه دهند. پدیدآورندگان اسناد علمی از تصاویر برای تشریح بهتر روش پیشنهادی خود یا برای نمایش نتایج خود و مقایسهٔ آن با روش‌های دیگر بهره می‌برند. به همین دلیل، اخیراً موتورهای جست‌وجو در کنار بازیابی اطلاعات از متن اسناد، به‌دنبال استخراج اطلاعات از تصاویر موجود در پایگاه اطلاعاتی خود نیز هستند.

«پایگاه اطلاعات علمی ایران (گنج)»، دستاوُرد کاربرد فناوری اطلاعات برای مدیریت اطلاعات علم و فناوری در «پژوهشگاه علوم و فناوری اطلاعات ایران (ایرانداک)» است. این پایگاه هم‌اکنون مرجع بسیاری از پژوهشگران ایران و جهان است و روزانه بیش از ده هزار کاربر، ده‌ها هزار جست‌وجو در آن انجام می‌دهند. این سامانه بنیان سامانه‌های دیگری در «ایرانداک» همچون «سامانهٔ همانندجو»، «سامانهٔ پیشینهٔ پژوهش»، و برخی از داشبوُردهای رصدخانهٔ پژوهش و فناوری نیز هست. هم‌اکنون در «گنج» امکان جست‌وجو بر اساس

یک عبارت متنی پرس‌وجو و بازیابی و نمایش نتایج جست‌وجو در قالب فراداده‌های متنی (عنوان، چکیده، کلیدواژه، پدیدآور، سال انتشار) وجود دارد. لیکن اطلاعات از تصاویر موجود در اسناد «گنج» بازیابی نمی‌شود. فراهم کردن چنین امکانی در «گنج» به‌عنوان یک ارزش افزوده می‌تواند آن را از موتورهای جست‌وجوی مشابه متمایز سازد.

بدین‌منظور، در این پژوهش قصد داریم روشی ساختارمحور برای ایجاد پایگاه داده از تصاویر مستخرج از اسناد علمی ارائه دهیم. در این راستا و در جهت بازیابی اطلاعات از تصاویر ابتدا باید تصاویر و توضیح متنی آن‌ها استخراج شود و پس از آن، تصاویر بر اساس محتوای آن‌ها طبقه‌بندی شده و برچسب بخورند. مجموعهٔ این اطلاعات به موتور جست‌وجو فرستاده می‌شود تا امکان دسترسی کاربر به تصویر مربوط به جست‌وجوی متنی به‌وجود بیاید. ادامهٔ این پژوهش به این ترتیب سازمان یافته است که ابتدا، در بخش دوم با مروری بر پیشینهٔ پژوهش، مهم‌ترین مطالعات صورت‌گرفته در این حوزه را بررسی خواهیم کرد. در بخش سوم، ضمن معرفی روش پژوهش، اجزای مختلف روش پیشنهادی را شرح خواهیم داد. برای بررسی کارایی روش پیشنهادی در بخش چهارم، از یک مطالعهٔ موردی در «پایگاه گنج» کمک خواهیم گرفت. سپس، نتایج حاصل از این روش را با روش پایه و مرسوم در ادبیات موضوع مقایسه خواهیم نمود. در نهایت، پژوهش در بخش آخر جمع‌بندی و نتیجه‌گیری می‌شود.

۲. پیشینهٔ پژوهش

امروزه، بیش از میلیون‌ها مقاله و سند علمی در وب وجود دارد (Khabsa and Giles 2014) و تعداد این اسناد به‌صورت نمایی در حال افزایش است؛ به‌گونه‌ای که جست‌وجو بین این حجم عظیم از اطلاعات علمی و فهمیدن نتایج و خلاصهٔ آن‌ها برای یک پژوهشگر به‌تنهایی تقریباً غیرممکن است (Milosevic et al. 2019). به همین دلیل، موتورهای جست‌وجو سعی می‌کنند از طریق الگوریتم‌های یادگیری ماشین، اطلاعات را از اسناد علمی موجود در پایگاه خود بازیابی و طبقه‌بندی کنند. با طبقه‌بندی اطلاعات می‌توان به کاربر برای جست‌وجوی مؤثر کمک کرد. تمرکز اکثر کارهایی که در حوزهٔ تحلیل اسناد انجام شده، بر روی متن اسناد است (علایی ابوذر ۱۳۹۷؛ فتاحی و نعیمی صدیق ۱۳۹۵؛ Chan, Ziftci and Forsyth 2006; Liu et al. 2007; Williams et al. 2014). الگوریتم‌های داده‌کاوی و

نمایه‌سازی موتورهای جست‌وجوی مقالات علمی، مثل «گوگل اسکالر»[1]، «سایت‌سیر»[2] و «گنج»[3] محدود به متن مقالات است و اطلاعات فرادادهٔ متنی مربوط به‌عبارت جست‌وجو را در اختیار قرار می‌دهند.

به‌دلیل اهمیتی که تصاویر در اسناد علمی پیدا کرده‌اند، روش‌های خودکار برای بازیابی اطلاعات از تصاویر نیز اخیراً مورد توجه قرار گرفته است (Futrelle et al. 2003; Chan, Ziftci, and Forsyth 2006; Xu, McCusker, and Krauthammer 2008; Li et al. 2013; Choudhury and Giles 2015; Tsutsi and Crandall 2017; Yang et al. 2017; Yu et al. 2017; Siegel et al. 2018).

برای بازیابی اطلاعات از اسناد علمی ابتدا باید بتوانیم نواحی گرافیکی و زیرنویس مربوط به آن‌ها را استخراج کنیم. روش استخراج نواحی گرافیکی می‌تواند بر اساس روش‌های پردازش تصویر باشد (Nagy and Seth 1984; Srihari 1986; Bloomberg 1991; Lemaitre, Camillerapp, and Coüasnon 2008; Cohen et al. 2013). در مطالعهٔ «ناگی و ست» تصویر حاصل از سند به بلوک‌هایی تقسیم می‌شود و با یک روش دسته‌بندی مناسب این بلوک‌ها به نواحی متن یا گرافیک تقسیم‌بندی می‌شوند (Nagy and Seth 1984). روش «بلومبرگ»، روش تشخیص متن از تصویر بر اساس تحلیل ریخت‌شناسی چندتفکیکی است که به‌دلیل پیاده‌سازی در نرم‌افزار «لپتونیکا»[4] به کرّات توسط کسانی که در پردازش تصاویر اسناد فعال هستند، مورد استفاده قرار می‌گیرد (Bloomberg 1991). «بلومبرگ» یک روش کاهش آستانه‌ای[5] را معرفی کرده است که بسته به مقدار آستانه می‌تواند به‌صورت روش‌های پایهٔ تحلیل ریخت‌شناسی به‌کار برود. با به‌کار بردن روش کاهش آستانه‌ای با آستانه‌های مختلف کلمات متن حذف می‌شوند و فیلتری دودویی ایجاد می‌شود که تنها در ناحیهٔ گرافیکی مقدار یک دارد. در این روش معمولاً تصاویری که محتوای متن آن‌ها زیاد است، به‌عنوان متن تقسیم‌بندی می‌شوند و بعضاً متن نزدیک تصویر به اشتباه تصویر تشخیص داده می‌شود.

به‌دلیل استقلال فایل «پی‌دی‌اف» از ویژگی‌های نرم‌افزاری و سخت‌افزاری سیستم، امروزه مقالات علمی بیشتر به‌صورت «پی‌دی‌اف» ثبت می‌شوند. به همین دلیل، روش‌های

1. https://scholar.google.com
2. https://citeseerx.ist.psu.edu
3. https://ganj.irandoc.ac.ir
4. http://www.leptonica.com
5. Threshold reduction

بازیابی اطلاعات از اسناد علمی در ادبیات بیشتر در مورد فایل «پی‌دی‌اف» است (Futrelle et al. 2003; Choudhury et al. 2013; Choudhury, Mitra, and Giles 2015; Clark and Divvala 2016). تصاویر موجود در فایل «پی‌دی‌اف» را می‌توان از خود فایل استخراج کرد. ابزارهایی برای استخراج تصاویر «پی‌دی‌اف» مثل Poppler[1] و PDFBox[2] وجود دارد. خروجی این ابزارها نتایج نامطلوب زیادی دارند و تصاویر بُرداری را نیز نمی‌توانند استخراج کنند.

اجزای فایل «پی‌دی‌اف» مثل متن و تصاویر، با استفاده از عملگرهای مختلف و مستقل ایجاد می‌شوند. به همین دلیل، استخراج همزمان تصاویر و زیرنویس مربوطه چالش‌برانگیز است. روش «کلارک» و «دیوالا» بر اساس این مشاهده است که معمولاً در مقالات علمی ناحیه‌ای که متن بدنه را دربرنمی‌گیرد و در مجاورت زیرنویس قرار دارد، ناحیه گرافیکی است (Clark and Divvala 2016). آن‌ها با استفاده از روش‌های ابتدایی پردازش تصویر کادرهای محصورکنندهٔ بلوک‌های متن و گرافیک را در هر صفحه تشخیص می‌دهند. بلوک‌های متن با استفاده از یک‌سری ویژگی‌های نگارشی به زیرنویس و متن بدنه تقسیم‌بندی می‌شود. در آخر، ناحیهٔ گرافیکی بزرگ در مجاورت بلوک زیرنویس به آن تخصیص داده می‌شود. این روش در مورد تصاویر چندگانه که از چند زیرتصویر تشکیل شده‌اند، درست عمل نمی‌کند.

«چادهاری» و همکاران تمام خطوطی را که شامل شناسهٔ تصویر هستند، استخراج می‌کنند (Choudhury et al. 2013). در این پژوهش، خطوط زیرنویس بر اساس ویژگی‌های از پیش تعیین‌شده مبتنی بر مشخصات ظاهری و نگارشی از سایر خطوط تشخیص داده می‌شوند. آن‌ها سپس، با استفاده از PDFBox تصاویر ماتریسی را استخراج می‌کنند. PDFBox موقعیت تصویر در فایل و مشخصات هندسی آن، مانند طول و عرض آن را فراهم می‌کند. آن‌ها با توجه به موقعیت تصویر یک مستطیل را زیرتصویر در نظر می‌گیرند و متن داخل آن را استخراج می‌کنند و از این متن استخراج‌شده شناسهٔ تصویر را به‌دست می‌آورند. شناسهٔ به‌دست‌آمده با شناسهٔ متناظر با هر زیرنویس مقایسه می‌شود و برای هر تصویر یک زیرنویس در نظر گرفته می‌شود. بعد از ایجاد پایگاه داده از تصاویر و توضیح متنی آن‌ها، این اطلاعات می‌توانند جهت بازیابی بهینه در موتور جست‌وجو پردازش شود. در این راستا «سیگل» و همکاران و «یو» و همکاران از روش‌های یادگیری عمیق برای

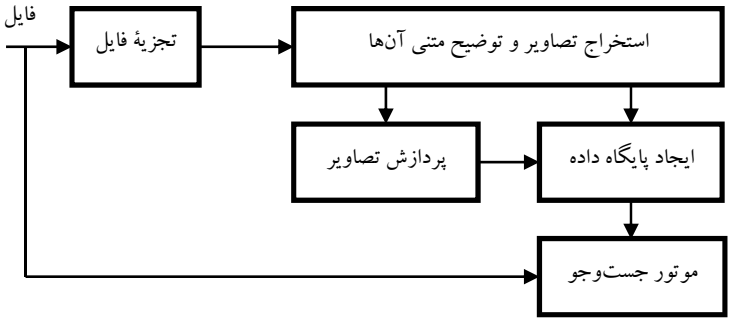
1. https://poppler.freedesktop.org 2. https://pdfbox.apache.org

گروه‌بندی تصاویر و استخراج اطلاعات از آن‌ها استفاده کرده‌اند (Siegel et al. 2016; Yu et al. 2017). در مطالعه‌ای دیگر «پیک ، ناکاگاوا و نوبل» برای کمک به انجام فراتحلیل نتایج علمی، اقدام به استخراج آمار توصیفی نظیر میانگین و انحراف معیار از تصاویر مربوط به چهار نوع نمودار خاص شامل نمودارهای میله‌ای، جعبه‌ای، پراکنده و هسیتوگرام کردند (Pick, Nakagawa and Noble 2019).

با نگاهی به مطالعات انجام‌شده در این حوزه درمی‌یابیم که هنوز روشی معرفی نشده است که بتواند بر مشکلات متعددی که بر سر راه استخراج تصاویر و متن مربوط به آن‌ها از اسناد علمی وجود دارد، فائق آید. این امر در کنار این واقعیت که استخراج متن از فایل‌های «پی‌دی‌اف» در پیشینهٔ موضوع، بیشتر متمرکز بر زبان‌هایی بوده است که از چپ به راست نگاشته می‌شوند، چالش دیگری را در پردازش فایل‌های «پی‌دی‌اف» موجود در زبان فارسی آشکار می‌سازد. در خروجی نرم‌افزارهای موجود برای استخراج متن «پی‌دی‌اف»، ترکیب کلمات جملات فارسی به‌هم می‌ریزد و جملات، معنادار نیستند. این مشکل در مورد جملاتی که ترکیب فارسی و لاتین دارند، بارزتر است. بدین‌ترتیب، تشخیص جملات و تصحیح ترکیب‌بندی کلمات در اسناد فارسی با فناوری‌های موجود چالش‌برانگیز است. از این رو، در ادامه سعی خواهیم کرد که برای غلبه بر این مشکلات روشی ساختارمحور مبتنی بر پردازش فایل وُرد برای استخراج تصاویر و توضیحات آن‌ها از اسناد علمی در زبان فارسی ارائه دهیم.

۳. روش پژوهش

در این پژوهش با استفاده از روش مطالعات اسنادی، سیستمی برای ایجاد پایگاه داده از تصاویر مستخرج از اسناد علمی طراحی شد. این پژوهش از نظر ماهیت در ردیف پژوهش‌های توسعه‌ای-کاربردی قرار می‌گیرد. شکل ۱، واحدهای مختلف سیستم طراحی‌شده را نشان می‌دهد. این سیستم با تجزیه و تحلیل فایل اسناد علمی کار خود را آغاز کرده و در نهایت، نتایج را برای بهره‌برداری‌های آتی در موتور جست‌وجو نمایه و آمادهٔ بازیابی می‌کند.



شکل ۱. مراحل بازیابی اطلاعات از تصاویر موجود در اسناد علمی

۳-۱. تجزیهٔ فایل

در بخش اول، فایل تجزیه شده و اطلاعات گرافیکی از متن جدا می‌شود. روش‌های استخراج تصویر از فایل را می‌توان به دو گروه تقسیم کرد. دستهٔ اول، مبتنی بر روش‌های پردازش تصویر و دستهٔ دوم مبتنی بر پردازش فایل سند است.

در رویکرد اول، صفحات سند علمی تبدیل به تصویر می‌شود و پس از آن از روش‌های موجود در پردازش تصویر برای ناحیه‌بندی تصویر صفحه استفاده می‌گردد تا اطلاعات گرافیکی از بدنهٔ اصلی متن جدا شود. این روش‌ها زیرمجموعه‌ای از روش‌های پردازش تصویر اسناد هستند که قدمت آن‌ها به حدود شصت سال پیش می‌رسد (Nagy and Seth 1984; Bloomberg 1991; Kumar et al. 2007; Lopez et al. 2011; Rehman and Saba 2014; Choudhury and Giles 2015; Choudhury, Mitra, and Giles 2015). امتیاز این روش‌ها آن است که در مورد تصاویر ماتریسی و برداری یکسان عمل می‌کنند و مستقل از فرمت فایل هستند. از معایب اکثر این روش‌ها آن است که فرض اولیهٔ آن‌ها بر این اصل استوار است که تراکم پیکسل‌های حاوی اطلاعات در متن سند علمی بیشتر از قسمت گرافیکی سند است. این فرض در مورد تصاویری که اطلاعات متنی زیادی دارند و یا مثلاً دربارهٔ تصاویر نقشه‌های جغرافیایی درست نیست. بنابراین، در اکثر مواقع برای گرفتن نتیجهٔ درست نیاز به تنظیم دستی پارامترهاست. سرعت این روش‌ها از آنجا که باید هر صفحه را به تصویر تبدیل کرده و الگوریتم پردازش تصویر را برای آن اجرا کنند، پایین است.

متن و تصاویر به‌صورت اشیای جداگانه در فایل «پی‌دی‌اف» و وُرد ذخیره می‌شوند. در روش‌های مبتنی بر پردازش فایل، فایل به اجزای تشکیل‌دهندهٔ آن تجزیه می‌شود (Futrelle

et al. 2003; Lopez et al. 2011; Choudhury et al. 2013; Praczyk and Nogueras-Iso 2013; Clark and Divvala 2016). این روش‌ها وابسته به فرمت و ساختار فایل بوده و برای فایل وُرد و «پی‌دی‌اف» متفاوت هستند. اکثر این روش‌ها در استخراج تصاویر بُرداری به مشکل برمی‌خورند. روش‌های مبتنی بر پردازش فایل، در مقایسه با روش‌های مبتنی بر پردازش تصویر، قابلیت مقیاس‌پذیری و سرعت بیشتری دارند و از این رو، برای به‌کارگیری در موتورهای جست‌وجو انتخاب بهتری هستند. استخراج زیرنویس مربوط به هر تصویر هم با توجه به فرمت فایل متفاوت است و در مورد فایل‌های «پی‌دی‌اف» بسیار چالش‌برانگیز است. از این رو، در این پژوهش از روش پردازش فایل وُرد برای استخراج تصاویر و توضیحات متنی کمک گرفته شده است.

۳-۲. استخراج تصاویر و توضیح متنی آن‌ها از فایل وُرد

برای استخراج تصویر و متن متناظر با آن نیازمند بهره‌برداری از ساختار فایل وُرد هستیم؛ از این رو روش پیشنهادی در این بخش یک روش ساختارمحور مبتنی بر چیدمان و آرایش فایل وُرد است. فایل وُرد با پسوند DOCX در واقع، یک فایل فشرده‌شده از مجموعه‌ای از فایل‌های XML است. اگر پسوند یک فایل DOCX را به zip تغییر داده و بعد آن را غیرفشرده[1] کنیم، مجموعه‌ای از فایل‌ها و فولدرها مطابق شکل ۲، تشکیل می‌شود. در ادامه، مهم‌ترین این فایل‌ها را معرفی خواهیم کرد.

شکل ۲. ساختار فولدربندی و فایل‌های XML مهم در یک فایل وُرد

1. unzip

فایل rels. یک مرجع[1] از محتوای فایل‌های سند را برای نرم‌افزارهای ویرایش متن نظیر «مایکروسافت وُرد»[2] فراهم می‌کند. مثلاً آدرس فایل اصلی XML را می‌توان در این فایل پیدا کرد. فایل document.xml.rels یک مرجع از منابع موجود در فایل وُرد مثل تصاویر جاسازی‌شده را فراهم می‌کند. فایل document.xml فایل XML اصلی است که متن سند و مشخصات ظاهری و طرح‌بندی هر صفحه و دستورالعمل‌های جای‌گذاری تصاویر و جداول را شامل می‌شود. تمام تصاویر جاسازی‌شده در سند در پوشهٔ مدیا[3] به‌طور جداگانه ذخیره می‌شوند.

چالش اصلی پیدا کردن زیرنویس مربوط به هر تصویر موجود در پوشهٔ مدیاست. جهت استخراج توضیح متنی تصاویر نیاز است فایل document.xml را تجزیه کنیم. برای تجزیه و تحلیل فایل XML روش‌های مختلفی وجود دارد. یکی از روش‌های متداول استفاده از زبان مسیر XML[4] یا به اختصار XPath است که اجازهٔ انتخاب و پیمایش بر روی گره‌های مختلف متناظر با ساختار درختی یک فایل XML را میسر می‌سازد.

متن هر سند از تعدادی پاراگراف و جدول تشکیل شده است (شکل ۳). هر پاراگراف که با گره <w:p> مشخص می‌شود، می‌تواند شامل متن سند و دستور جای‌گذاری تصویر باشد. هر جدول که با گره <w:tbl> نشان داده می‌شود، خود می‌تواند شامل پاراگراف و یا جدول دیگری باشد. متن که در واقع، مجموعه‌ای از کاراکترهاست با گره <w:t> مشخص می‌شود. حدود چهل برچسب مختلف ویژگی‌های ظاهری متن را مثل فونت، سایز، رنگ و قلم مشخص می‌کنند. برای آشنایی بیشتر با این برچسب‌ها به (Office Open XML 2019) مراجعه کنید.

1. reference
2. Microsoft Word
3. media
4. XML Path Language

```
<w:document>
    <w:body>
        <w:p>
            <w:pPr/>
            <w:r>
                <w:t>
                </w:t>
            </w:r>
        </w:p>
        …
        <w:tbl>
            <w:tblPr/>
            <w:tblGrid/>
            <w:tr>
                <w:tc>
                    <w:tcPr/>
                    <w:p>
                        <w:r>
                            <w:t>
                            </w:t>
                        </w:r>
                    </w:p>
                </w:tc>
                <w:tc>
                    …
                </w:tc>
                …
            </w:tr>
            …
        </w:tbl>

        …
    </w:body>
</w:document>
```

شکل ۳. ساختار کلی فایل document.xml

فایل DOCX دو نوع تصویر درون‌خطی و شناور را پشتیبانی می‌کند. تمام اطلاعات مربوط به سایز و طرح‌بندی تصاویر در گره <w:drawing> می‌آید. هر تصویردرون‌خطی در فایل XML یک شناسه دارد و فایل document.xml.rels مشخص می‌کند که این شناسه مربوط به کدام تصویر استخراج‌شده در پوشهٔ مدیاست. تصاویر شناور مانند کاراکتر متن در نظر گرفته می‌شوند و متن در اطراف آن‌ها جاری می‌شود. تصاویر شناور با گره <wp:anchor> در درون گره <w:drawing> قرار داده می‌شوند.

استخراج زیرنویس با جست‌وجو در پاراگراف‌های بعد از پاراگراف مربوط به تصویر انجام می‌شود. در بسیاری از موارد، زیرنویس تصاویر با استفاده از قابلیت درج زیرنویس نرم‌افزار «مایکروسافت وُرد»، نوشته می‌شود. در این موارد متن زیرنویس در عنصر <caption> بعد از تصویر ذخیره می‌شود. اگر از قابلیت درج زیرنویس استفاده نشده باشد، استخراج زیرنویس پیچیده‌تر می‌شود.

با توجه به آنچه که گفته شد، در هر سند به‌دنبال تصاویری هستیم که درون پاراگراف‌ها قرار گرفته‌اند. بدین‌ترتیب، برای هر پاراگرافِ شامل تصویر (پاراگراف هدف)، دو حالت کلی وجود دارد: این پاراگراف یا در بدنهٔ اصلی سند قرار دارد و یا این‌که داخل یک جدول است. در صورتی که پاراگراف هدف در بدنهٔ اصلی سند باشد، برای استخراج زیرنویس متناظر با آن تنها کافی است که عناصر بعد از این پاراگراف را برای عنصر <caption> جست‌وجو نماییم. در حالتی که پاراگراف درون یکی از سلول‌های جدول (سلول هدف) باشد، سه مکان مختلف باید برای استخراج زیرنویس تصویر جست‌وجو شود. ابتدا باید زیرنویس را در عناصری بررسی کنیم که بعد از پاراگراف هدف و در داخل همان سلول از جدول درج شده‌اند (زیرنویس درون سلولی). سپس، باید زیرنویس را در عناصر سلول‌هایی بررسی نماییم که در سطر بعد و در زیر سلول هدف قرار دارند (زیرنویس بین سلولی). در نهایت، زیرنویس را باید در عناصر بعد از جدول هدف جست‌وجو کنیم (زیرنویس بعد از جدولی). گام‌های الگوریتم ساختارمحور پیشنهادی در ادامه تشریح شده است.

گام ۱: فایل وُرد را بارگذاری و غیرفشرده کن.

گام ۲: فایل document.xml را بارگذاری کرده و ساختار درختی متناظر با آن را استخراج کن.

گام ۳: برای هر پاراگراف در ساختار درختی فوق موارد زیر را انجام بده:

گام ۳-۱: گره تصویر را پیدا کن.

گام ۳-۲: اگر پاراگراف دربرگیرندهٔ گره تصویر (پاراگراف هدف) داخل یک جدول است، به گام ۳-۳ برو. در غیر این صورت به گام ۳-۵ برو.

گام ۳-۳: برای تمام جداولی که پاراگراف هدف را دربرگرفته‌اند، موارد زیر را انجام بده:

گام ۳-۳-۱: برای تمام عناصر همرده (siblings) پاراگراف هدف موارد زیر را انجام بده:

گام ۳-۳-۱-۱: اگر عنصر بعدی یک پاراگراف بدون عکس و دارای متن بود، متن آن را به‌عنوان برچسب تصویر استخراج کن و به گام ۳-۳-۲ برو.

گام ۳-۳-۱-۲: اگر عنصر بعدی جدول بود، برای سطرهای آن، موارد زیر را انجام بده:

گام ۳-۳-۱-۲-۱: اگر سطر بدون عکس و دارای متن بود، متن آن را به‌عنوان برچسب تصویر استخراج کن و به گام ۳-۳-۲ برو.

گام ۳-۳-۱-۲-۲: اگر سطر دارای عکس بود، متن مربوط به این سطر را به‌عنوان برچسب تصویر استخراج کن و به گام ۳-۳-۲ برو.

گام ۳-۳-۱-۳: اگر عنصر بعدی، پاراگراف و یا جدول نبود، به گام ۳-۳-۲ برو.

گام ۳-۳-۲: سطر و سلول دربرگیرندهٔ پاراگراف هدف (سطر و سلول هدف) را پیدا کن.

گام ۳-۳-۳: برای تمام سلول‌هایی که در سطر بعد از سطر هدف و در زیر سلول هدف قرار دارند، موارد زیر را انجام بده:

گام ۳-۳-۳-۱: اگر سلول بدون عکس و دارای متن بود، متن آن را به‌عنوان برچسب تصویر استخراج کن و به گام ۳-۳-۴ برو.

گام ۳-۳-۳-۲: اگر سلول بدون عکس و متن بود، ابتدا سلول‌هایی را که

در سطری بعدی زیر آن قرار دارند، پیدا کرده و سپس، متن آن‌ها را به‌عنوان برچسب تصویر استخراج کن و به گام ۳-۳-۴ برو.

گام ۳-۳-۳-۳: اگر سلول دارای عکس و متن بود، متن آن را به‌عنوان برچسب تصویر استخراج کن و به گام ۳-۳-۴ برو.

گام ۳-۳-۳-۴: اگر سلول دارای عکس و بدون متن بود، ابتدا سلول‌هایی را که در سطری بعدی زیر آن قرار دارند، پیدا کرده و سپس، متن آن‌ها را به‌عنوان برچسب تصویر استخراج کن و به گام ۳-۳-۴ برو.

گام ۳-۳-۴: برای عناصر همردهٔ جدول دربرگیرندهٔ پاراگراف هدف (جدول هدف) موارد زیر را انجام بده:

گام ۳-۳-۴-۱: اگر عنصر بعدی یک پاراگراف بدون عکس و دارای متن بود، آنگاه اگر عنصر <caption> داشت، متن آن را به‌عنوان برچسب تصویر استخراج کن. در نهایت، به گام ۳-۴ برو.

گام ۳-۳-۴-۲: اگر عنصر بعدی جدول بود، برای سطرهای آن موارد زیر را انجام بده:

گام ۳-۳-۴-۲-۱: اگر سطر بدون عکس و دارای متن بود، آنگاه اگر عنصر <caption> داشت، متن آن را به‌عنوان برچسب تصویر استخراج کن. در نهایت، به گام ۳-۴ برو.

گام ۳-۳-۴-۲-۲: اگر سطر دارای عکس بود، آنگاه اگر عنصر <caption> داشت، متن مربوط به این سطر را به‌عنوان برچسب تصویر استخراج کن. در نهایت، به گام ۳-۴ برو.

گام ۳-۳-۴-۳: اگر عنصر بعدی پاراگراف و یا جدول نبود به گام ۳-۴ برو.

گام ۳-۴: اگر حداقل یکی از متن‌های استخراج‌شدهٔ فوق عنصر <caption> داشت، تصویر و توضیحات متناظر با آن را ذخیره کن و به گام ۳-۶ برو.

گام ۳-۵: برای عناصر هم‌ردهٔ پاراگراف هدف موارد زیر را انجام بده:

گام ۳-۵-۱: اگر عنصر بعدی یک پاراگراف بدون عکس و دارای متن بود، آنگاه اگر عنصر <caption> داشت، متن آن را به‌عنوان برچسب تصویر

استخراج کرده و سپس، تصویر و متن متناظر را ذخیره کن. در نهایت، به گام ۳-۶ برو.

گام ۳-۵-۲: اگر عنصر بعدی جدول بود، برای سطرهای آن موارد زیر را انجام بده:

گام ۳-۵-۲-۱: اگر سطر، بدون عکس و دارای متن بود، آنگاه اگر عنصر <caption> داشت، متن آن را به‌عنوان برچسب تصویر استخراج کرده و سپس، تصویر و متن متناظر را ذخیره کن. در نهایت، به گام ۳-۶ برو.

گام ۳-۵-۲-۲: اگر سطر دارای عکس بود، آنگاه اگر عنصر <caption> داشت، متن مربوط به این سطر را به‌عنوان برچسب تصویر استخراج کرده و سپس، تصویر و متن متناظر را ذخیره کن. در نهایت، به گام ۳-۶ برو.

گام ۳-۵-۳: اگر عنصر بعدی، پاراگراف و جدول نبود، به گام ۳-۶ برو.

گام ۳-۶. اگر گره، تصویر دیگری ندارد، توقف کن، در غیر این صورت به گام ۳-۱ برو.

۳-۳. پردازش تصویر

استخراج اطلاعات از خود تصاویر نیز می‌تواند به بازیابی بهتر تصاویر کمک کند. تصاویر در اسناد علمی شامل نمودار میله‌ای، نمودار پای، نمودار خطی، تصاویر طبیعی و امثال آن می‌شوند. نوع تصویر می‌تواند به‌عنوان یک برچسب برای تصویر در نظر گرفته شود. روش‌های متعددی برای گروه‌بندی تصاویر استخراج‌شده از اسناد علمی وجود دارد. در مطالعهٔ «ساوا» و همکاران از ایدهٔ سبد کلمه‌های تصویری برای استخراج ویژگی‌های هر تصویر استفاده شد و نویسندگان با استفاده از روش گروه‌بندی ماشین بردار پشتیبان، تصاویر را طبقه‌بندی کردند (Savva et al. 2011). «سیگل» و همکاران از روش شبکه‌های عصبی پیچشی[1] استفاده کردند (Siegel et al. 2016).

متن موجود در تصاویر هم می‌تواند به درک بهتر تصویر کمک کند. استخراج متن

1. convolutional neural networks

از تصویر با استفاده از روش‌های نویسه‌خوان نوری[1] انجام می‌شود. امروزه، نرم‌افزارهای تجاری زیادی وجود دارند که به‌خوبی می‌توانند متن انگلیسی موجود در تصویر به همراه موقعیت آن را استخراج کنند. نرم‌افزارهای نویسه‌خوان نوری مایکروسافت (Microsoft 2019)، گوگل (Smith 2007) و اَبی (ABBYY 2019) از آن جمله هستند. این نرم‌افزارها دقت پایین‌تری در مورد متون فارسی دارند.

۳-٤. ایجاد پایگاه داده تصاویر

بعد از آن‌که تصاویر و اطلاعات مربوطه از فایل استخراج شد، باید آن‌ها را مطابق با یک مدل داده‌ای که توسط موتور جست‌وجوگر قابل استفاده باشد، ذخیره‌سازی کنیم. در جدول ۱، مدل دادهٔ پیشنهادی ارائه شده است. توضیحات این جدول به شرح زیر است:

۱. id: شناسهٔ اصلی هر تصویر در پایگاه داده است که نوعاً به‌صورت ترتیبی افزایش می‌یابد.
۲. article-id: شناسهٔ سندی که تصویر استخراج شده از آن را نشان می‌دهد. این شناسه در نقش یک کلید خارجی به جدول اسناد است.
۳. uid: شناسهٔ کدشده است که یک رشتهٔ ۳۲ بیتی است. به‌منظور نگهداری امن تصاویر استخراج‌شده لازم است، نامی که برای هر تصویر انتخاب می‌شود، معادل با شناسهٔ کدشدهٔ آن باشد.
۴. title_fa: زیرنویس فارسی تصویر است.
۵. title_en: چنانچه زیرنویس تصویر انگلیسی باشد، در این قسمت ذخیره می‌شود.

در ردیف‌های ۶ الی ۸ مشخصات ظاهری تصویر ذخیره می‌شوند. موارد ۹ و ۱۰ تاریخ ورود و آخرین به‌روزرسانی داده در پایگاه اطلاعاتی است.

1. Optical Character Reader (OCR)

جدول ۱. مدل داده برای ایجاد پایگاه اطلاعاتی تصاویر

ردیف	مشخصه	نوع داده‌ای	توضیح
۱	id	integer	کلید اصلی
۲	article_id	integer	شناسهٔ مدرک، کلید خارجی به جدول اسناد (articles)
۳	uid	string	شناسهٔ کدشده جدول
۴	title_fa	string	عنوان فارسی تصویر
۵	title_en	string	عنوان انگلیسی تصویر
۶	size	integer	اندازهٔ تصویر به کیلوبایت
۷	height	integer	ارتفاع تصویر به پیکسل
۸	width	integer	عرض تصویر به پیکسل
۹	created_at	datetime	تاریخ ایجاد رکوُرد
۱۰	updated_at	datetime	تاریخ آخرین به‌روزرسانی رکوُرد

شکل ۴، نمودار رابطهٔ موجودیت‌ها (ERD)[1] را برای پایگاه دادهٔ تصاویر نشان می‌دهد. در این شکل، ارتباط بین جدول مربوط به تصاویر (images) با دیگر جداول مربوط به اسناد علمی قابل مشاهده است. مطابق آنچه دیده می‌شود، رابطهٔ بین جدول اسناد (articles) با جدول images یک‌به‌چند[2] است؛ بدین‌معنا که برای هر سند چند تصویر می‌تواند وجود داشته باشد. رابطهٔ بین جدول articles با جداول سازمان‌ها (organizations)، کلیدواژه‌ها (keywords)، و پدیدآوران (authors) چندبه‌چند است؛ یعنی به‌طور مثال، هر سند می‌تواند چند کلیدواژه و هر کلیدواژه می‌تواند در چندین سند وجود داشته باشد. کلیدهای خارجی با رنگ نارنجی در این شکل متمایز شده‌اند و وظیفهٔ آن‌ها ایجاد ارتباط بین جدول‌های مختلف است.

1. entity relationship diagram
2. one to many

شکل ٤. نمودار رابطهٔ موجودیت‌ها برای پایگاه اطلاعاتی تصاویر

۳-۵. موتور جست‌وجو

در این واحد، اطلاعات ذخیره‌شده در پایگاه داده ایجاد شده، در موتور جست‌وجو نمایه شده و جهت بازیابی اطلاعات مورد استفاده قرار خواهد گرفت. لازم به ذکر است که می‌توان نمایه‌سازی را از جهات مختلفی نظیر محتوای اطلاعاتی برچسب تصاویر، مشخصات ظاهری تصاویر و اطلاعات مربوط به برچسب‌های حاصل از پردازش تصاویر

انجام داد. اهمیت استفاده از هر یک از این دسته دادگان بسته به نوع کاربری مورد انتظار از بازیابی تصاویر نمایه‌شده متفاوت است.

٤. مطالعهٔ موردی

برای بررسی کارایی روش پیشنهادی در استخراج تصاویر و توضیحات آن‌ها از یک مطالعهٔ موردی در «پایگاه اطلاعات علمی ایران (گنج)» کمک گرفتیم. این پایگاه مرجع اصلی دسترسی به تمام متن پایان‌نامه‌ها و رساله‌های تحصیلات تکمیلی در ایران است. به‌صورت تصادفی ۱۵۰ سند فنی-مهندسی را از پایگاه دادهٔ «گنج» انتخاب کردیم که اکثر این اسناد پایان‌نامه‌های ارشد و دکتری هستند. در ادامه، این اسناد توسط روش پیشنهادی مورد تجزیه و تحلیل قرار گرفتند. روش پیشنهادی از فایل وُرد استفاده می‌کند. برای ارزیابی عملکرد روش پیشنهادی، آن را با روش پایهٔ مرسوم در پیشینهٔ موضوع مقایسه کردیم. روش‌های پایه از فایل «پی‌دی‌اف» برای استخراج تصاویر و متن آن‌ها استفاده می‌کنند.

٤-١. پیاده‌سازی روش پیشنهادی

برای پیاده‌سازی روش پیشنهادی از زبان برنامه‌نویسی «پایتون»[1] و «کتابخانهٔ lxml» کمک گرفته شد. این کتابخانه فایل XML را به‌صورت درختی از عنصرها بارگذاری می‌کند. عنصر، شیء اصلی در این کتابخانه است که به‌طور اختصاصی برای ذخیره‌سازی داده‌های دارای سلسله‌مراتب طراحی شده است. در حقیقت، عنصر حد واسط بین نوع داده‌ای لیست و دیکشنری در زبان «پایتون» است. فایل بارگذاری‌شده توسط درخت عنصرها قابلیت جست‌وجو و استخراج داده با دستورالعمل‌های «پایتون» را پیدا می‌کند. تمام اطلاعات مربوط به طرح‌بندی تصویر در عنصر <drawing> ذخیره می‌شود. بدین‌ترتیب، شناسهٔ هر تصویر با استفاده از Xpath زیر قابل بازیابی است.

w:drawing/wp:inline/a:graphic/a:graphicData/pic:pic/pic:blipFill/a:blip/@r:embed

شناسهٔ هر تصویر مشخص می‌کند که اطلاعات ذخیره‌شده در فایل XML به کدام تصویر استخراج‌شده در پوشهٔ مدیا مربوط می‌شود. برای پیدا کردن زیرنویس مربوط به هر تصویر باید پاراگراف‌ها و جداول بعد از دستور جای‌گذاری تصویر را بررسی کرد. در

1. Python

مورد فایل‌هایی که در نگارش از قابلیت درج زیرنویس وُرد استفاده شده است، اطلاعات مربوط به زیرنویس در عنصر <caption> ذخیره می‌شود. تصاویر و زیرنویس مربوط به آن‌ها مطابق الگوریتم ارائه‌شده در بخش ۳-۲ استخراج می‌شوند.

٤-۲. پیاده‌سازی روش پایه

همان‌طور که پیش‌تر گفته شد، روش ایجاد پایگاه اطلاعات تصاویر در ادبیات در مورد فایل‌های «پی‌دی‌اف» است. در فایل «پی‌دی‌اف» اطلاعات در زیر شیء صفحه ذخیره می‌شود. تمام اطلاعات مهم برای ایجاد یک صفحهٔ «پی‌دی‌اف» در این شیء آورده شده است. «پی‌دی‌اف» معمولاً تصاویر ماتریسی را در یک شیء جداگانه به نام Xobject به‌صورت دادهٔ دودویی ذخیره می‌کند. برای استخراج تصویر باید همهٔ اطلاعات تصویر را که در Xobject ذخیره شده، استخراج کرد و تصویر را با استفاده از آن ایجاد نمود.

بازیابی تصاویر بُرداری بسیار پیچیده‌تر از تصاویر ماتریسی است. تصاویر بُرداری در قالب مسیر می‌آید که در واقع، رشته‌ای از کاراکترها هستند. به‌عنوان مثال، در مسیر زیر دستور ایجاد یک خط در فایل «پی‌دی‌اف» آمده است.

175 720 m 175 50 m l h S

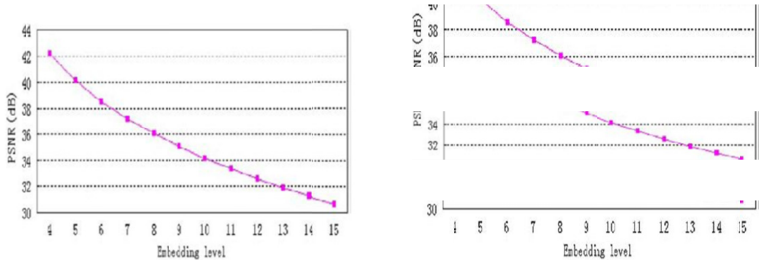
اعداد قبل از m نقطهٔ شروع مسیر را نشان می‌دهد، l تعیین می‌کند که نوع مسیر خط راست است، و h نشان می‌دهد که 175 50 نقطهٔ انتهای مسیر است و S دستور کشیدن خط است. یک تصویر از مجموعه‌ای از مسیرها ساخته می‌شود. گاهی اوقات متن موجود در «پی‌دی‌اف» هم به‌صورت بُردار گرافیکی ایجاد شده و با مسیر تعیین می‌شود. تشخیص درست مسیرهای مربوط به یک تصویر بسیار چالش‌برانگیز است.

برای استخراج تصویر از فایل «پی‌دی‌اف»، کتابخانه‌هایی وجود دارد که Apache PDFBox، XPDF، PyPDF2 و Poppler از آن جمله هستند. این کتابخانه‌ها منبع باز و رایگان هستند. PyPDF با زبان «پایتون» و Apache PDFBox با زبان «جاوا»[1] پیاده‌سازی شده است. کتابخانه‌ها به‌صورتی که وجود دارند، فقط برای استخراج تصاویر ماتریسی مناسب هستند و قادر به استخراج تصاویر بُرداری نیستند.

برای پردازش فایل‌های «پی‌دی‌اف» آزمایشی از کتابخانهٔ PDFBox در «جاوا» استفاده

1. Java

کردیم. PDFBox کتابخانه‌ای قدرتمند است که در مقالات پژوهشی مختلفی برای تجزیهٔ فایل «پی‌دی‌اف» مورد استفاده قرار گرفته است (Choudhury et al. 2013; Choudhury and Giles 2015; Clark and Divvala 2016; Siegel et al. 2016). کلاس getXObject شیء XObject را استخراج می‌کند که شامل اطلاعات مربوط به تصویر است. کلاس PDImageXObject با استفاده از اطلاعات XObject تصویر مورد نظر را ایجاد می‌کند. PDFBox تصاویر بُرداری را که با مسیر مشخص می‌شوند، استخراج نمی‌کند. در بعضی موارد، فلوچارت‌ها، جدول‌ها و فرمول‌ها از ترکیب چند XObject کوچک تشکیل می‌شوند. مشخصات هر خط و هر علامتی که در جدول وجود دارد، در یک XObject ذخیره شده و به‌صورت مجزا در خروجی نمایش داده می‌شود. این امر منجر به استخراج ده‌ها هزار تصویر نامطلوب می‌گردد. خیلی از این تصاویر نامطلوب، سایز کوچکی (حدود چند ده پیکسل) دارند. به همین دلیل، تنها تصاویری که ارتفاع و عرض آن‌ها از ۵۰ پیکسل بیشتر باشند، در این آزمایش مد نظر قرار گرفته‌اند. با این فیلتر تعداد زیادی از تصاویر نامطلوب حذف می‌شود، اما همچنان صدها تصویر نامطلوب باقی می‌مانند که از نظر ویژگی‌های ظاهری و ساختاری شباهت زیادی به تصاویر هدف دارند. همچنین، در بعضی موارد یک نمودار واحد ممکن است در تعدادی XObject ذخیره شود. مثلاً نمودار موجود در شکل ۵، از چند XObject جدا تشکیل شده است و در خروجی هر قسمت نمودار به‌صورت مجزا نمایش داده می‌شود.



شکل ۵. سمت چپ، نمودار نمایش داده‌شده در فایل «پی‌دی‌اف» و سمت راست، خروجی حاصل از نرم‌افزار تجزیهٔ فایل

متن در «پی‌دی‌اف» با استفاده از عملگرهای متفاوت از تصویر و به‌صورت مستقل ایجاد می‌شود. جهت استخراج زیرنویس در فایل‌هایی که متن آن‌ها تصاویر بُرداری نیستند، ابتدا جملات شامل کلمات کلیدی مثل «شکل» و «نمودار» استخراج می‌شوند. سپس، جملات

نامربوط بر اساس یک بُردار ویژگی‌های از پیش تعریف‌شده حذف می‌گردند. لازم به ذکر است که متن موجود در فایل‌های آزمایشی را با استفاده از کلاس getTex استخراج کردیم.

٤-٣. تجزیه و تحلیل نتایج

با بررسی فایل‌های وُرد دیده می‌شود که از ۱۵۰ فایل، ۱۳ سند وُرد از نظر محتوایی ناقص بودند. فایل وُردِ ۱۳۷ سندِ باقی‌مانده با برنامهٔ توسعه‌داده‌شده مورد تجزیه و تحلیل قرار گرفت. از ۱۳۷ فایل آزمایشی در ۵۵ سند از قابلیت درج زیرنویس استفاده شده است. اطلاعات مربوط به تصاویر در این ۵۵ سند مطابق جدول ۱، در پایگاه داده ذخیره‌سازی می‌شود. نتایج نشان می‌دهد که روش پیشنهادی قابلیت آن را دارد که حدود ۴۰ درصد از تصاویر را به همراه توضیحات مربوط به آن‌ها بدون خطا بازیابی نماید.

روش پایه تنها بر روی فایل‌های «پی‌دی‌اف» کاربرد دارد. از ۱۳۷ فایل «پی‌دی‌اف» در دادهٔ آزمایشی، متن ۵۹ فایل در واقع، تصویر بُرداری هستند و نمی‌توان آن‌ها را استخراج کرد. تقریباً در نیمی از فایل‌های باقی‌مانده، به‌خاطر فرمت خاص «پی‌دی‌اف» استفاده‌شده صدها تصویر نامطلوب ایجاد می‌شود که به‌دلیل شباهت ساختاری زیاد آن‌ها با تصاویر هدف، فیلترکردن آن‌ها بسیار چالش‌برانگیز است. جملات استخراج‌شده از فایل‌های «پی‌دی‌اف» فارسی ترکیب چپ به راست دارند و آرایش جملات فارسی خروجی، در جایی که با کلمات انگلیسی، علایم و یا فرمول ترکیب شده‌اند، به‌هم می‌ریزد. در بهترین سناریو، چنانچه بر مشکلات حذف تصاویر نامطلوب و تخصیص زیرنویس فائق بیاییم، حدوداً ۳۰ درصد تصاویر با زیرنویس آن‌ها را می‌توانیم با روش پایه از فایل‌های «پی‌دی‌اف» استخراج کنیم. بنابراین، دیده می‌شود که روش پایهٔ موجود برای استخراج تصویر و توضیح متنی آن در مورد فایل‌های فارسی کارایی لازم را ندارد و روش پیشنهادی به‌طور معناداری عملکرد بهتری را از خود نشان می‌دهد.

۵. نتیجه‌گیری

در این مقاله راه‌های استخراج تصویر و زیرنویس جهت ایجاد پایگاه دادهٔ تصاویر از اسناد علمی بررسی شد. در این راستا دو فرمت رایج «وُرد» و «پی‌دی‌اف» مورد مطالعه و مقایسه قرار گرفتند. سپس، با توجه به نوع نیازمندی‌های اسناد علمی منتشرشده به

زبان فارسی، روشی ساختارمحور و مبتنی بر پردازش فایل وُرد برای ایجاد پایگاه دادهٔ تصاویر پیشنهاد شد. در ادامه، عملکرد این روش با روش پایهٔ پردازش فایل «پی‌دی‌اف» در یک مطالعهٔ موردی در «پایگاه اطلاعات علمی ایران (گنج)» مورد مقایسه قرار گرفت. اسناد علمی در «گنج» به دو صورت «پی‌دی‌اف» و «وُرد» ذخیره می‌شوند که بخش بزرگی از آن‌ها را تمام‌متن پایان‌نامه‌ها و رساله‌ها تشکیل می‌دهد. نتایج تجربی، کارایی روش پیشنهادی را در مقابل روش پایه نشان داد. از طرف دیگر، دیده شد که در مورد فایل «پی‌دی‌اف» در زبان فارسی چالش‌های زیادی وجود دارد. تصاویر بُرداری را نمی‌توان استخراج کرد. تصاویر ماتریسی هم بعضاً به‌صورت اجزای جداگانه استخراج می‌شوند. از آنجا که متن موجود در سند تصویر بُرداری است، حدوداً در یک سوم فایل‌ها قادر به استخراج متن برای پیدا کردن زیرنویس نخواهیم بود. به‌دلیل این مشکلات پیشنهاد می‌کنیم برای ایجاد پایگاه دادهٔ تصاویر از روش پیشنهادی و نسخه وُرد فایل‌ها استفاده شود.

برای مطالعه‌های آتی می‌توان الگوریتم پیشنهادی را بهبود بخشید. برخی از فایل‌ها اگرچه فرمت استاندارد ندارند، اما زیرنویس آن‌ها به شکل شیء مجزا در نرم‌افزار «مایکروسافت وُرد» قابل تشخیص است. در مورد این فایل‌ها می‌توان یک نمونهٔ آماری معتبر جمع‌آوری کرد و چنانچه بین همهٔ آن‌ها از نظر ساختاری شباهت وجود داشته باشد، می‌توان یک الگوریتم قانون‌محور برای استخراج زیرنویس آن‌ها طراحی نمود. جهت بهبود عملکرد جست‌وجو نیز می‌توان یک جدول برچسب برای تصاویر ایجاد کرد. این جدول برای هر تصویر حاوی برچسب‌هایی است که با پردازش تمام جملاتی حاصل می‌شود که در آن به تصویر مد نظر اشاره شده و کلمات کلیدی‌شان با استفاده از روش‌های پردازش زبان طبیعی استخراج شده است. همچنین، می‌توان با استفاده از روش‌های پردازش تصویر و نویسه‌خوانی، متن داخل تصویر را هم استخراج کرد و به‌عنوان برچسب در نظر گرفت. از طرف دیگر، کلیهٔ تصاویر استخراج‌شده را می‌توان به یکی از گروه‌های نمودار خطی، فلوچارت، تصاویر طبیعی، جدول و غیره تقسیم‌بندی کرد و گروه تصویر هم به‌عنوان یک برچسب در نظر گرفته شود. برای این کار می‌توان از روش‌های یادگیری عمیق بهره گرفت.

٦. قدردانی

این مقاله مستخرج از طرح پژوهشی است که با حمایت‌های مادی و معنوی «پژوهشگاه علوم و فناوری اطلاعات ایران (ایرانداک)» به انجام رسیده است. همچنین، نویسندگان از «مرکز فناوری اطلاعات ایرانداک» جهت فراهم‌کردن دادهٔ آزمایشی برای این پژوهش و از آقایان «مهندس مجتبی زالی» و «مهندس منصور شیدایی» برای کمک‌های مشورتی در مورد ساختار پایگاه دادهٔ «گنج» قدردانی می‌کنند.

فهرست منابع

علایی ابوذر، الهام. ۱۳۹۷. معرفی رویکردی ماشینی با استفاده از الگوریتم لسک و برچسب‌دهی نحوی جهت رفع ابهام از معنای کلمات. *پژوهشنامه پردازش و مدیریت اطلاعات* ۳۳ (۳): ۱۱۶۵-۱۱۸۲.

فتاحی، سمیه، و علی نعیمی صدیق. ۱۳۹۵. تحلیل رفتار اطلاع‌یابی پژوهشگران در موتور جست‌وجوی سامانه ملی اطلاعات پایان‌نامه‌ها/ رساله‌های دانش‌آموختگان داخل کشور (گنج). *نشریه علمی مدیریت اطلاعات* ۲ (۲): ۳۱-۵۸.

References

ABBYY. ABBYY FineReader 14. *PDF Software with Text Recognition*. https://www.abbyy.com/en-ee/finereader/ (accessed March 11, 2019).

Bloomberg, Dan S. 1991. *Multiresolution morphological approach to document image analysis*. In Proceedings of the International Conference on Document Analysis and Recognition. Boston, MA. pp. 963-971.

Chan, J., C. Ziftci, and D. Forsyth. 2006. *Searching off-line Arabic documents*. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognitio. New York, NY, USA. pp. 1455-1462.

Choudhury, S. R., and C. L. Giles. 2015. *An architecture for information extraction from figures in digital libraries*. In Proceedings of the 24th International Conference on World Wide Web Companion. Florance, Italy. pp. 667-672.

_____, P. Mitra, A. Kirk, S. Szep, D. Pellegrino, S. Jones, and C. L. Giles. 2013. *Figure metadata extraction from digital documents*. In Proceedings of the 12th International Conference on Document Analysis and Recognition. Washington, DC, USA. pp. 135-139.

_____, P. Mitra, and C. L. Giles. 2015. *Automatic extraction of figures from scholarly documents*. In Proceedings of the 2015 ACM Symposium on Document Engineering. New Yor, NY, USA. pp. 47-50.

Clark, C., and S. Divvala. 2016. *PDFFigures 2.0: Mining figures from research papers*. In Proceedings of the 16th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL). Newark, NJ. USA. pp. 143-152.

Cohen, R., A. Asi, K. Kedem, J. El-Sana, and I. Dinstein. 2013. *Robust text and drawing segmentation algorithm for historical documents*. In Proceedings of the 2nd International Workshop on Historical Document Imaging and Processing. New York, NY, USA. pp. 110-117.

Futrelle, R. P., M. Shao, C. Cieslik, and A. E. Grimes. 2003. *Extraction, layout analysis and classification of diagrams in PDF documents*. In Proceedings of the 7th International Conference on Document

Analysis and Recognition. Washington, D. C, USA. pp. 1007-1013.

Khabsa, M., and C. L. Giles. 2014. The number of scholarly documents on the public web. *PLoS ONE* 9 (5): 1-6.

Kumar, S., R. Gupta, N. Khanna, S. Chaudhury, and S. D. Joshi. 2007. Text extraction and document image segmentation using matched Wavelets and MRF model. *IEEE Transactions on Image Processing* 16 (8): 2117-2128.

Lemaitre, A., J. Camillerapp, and B. Coüasnon. 2008. Multiresolution cooperation makes easier document structure recognition. *International Journal of Document Analysis and Recognition* 11 (2): 97-109.

Li, Z., M. Stagitis, S. Carberry, and K. F. McCoy. 2013. *Towards retrieving relevant information graphics*. In Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval. Dublin, Ireland. pp. 789-792.

Liu, Y., K. Bai, P. Mitra, and C. L. Giles. 2007. *TableSeer: Automatic table metadata extraction and searching in digital libraries*. In Proceedings of the 7th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL). Vancouver, B.C. Canada. pp. 91-100.

Lopez, L., J. Yu, C. N. Arighi, H. Huang, H. Shatkay, and C. Wu. 2011. *An automatic system for extracting figures and captions in biomedical PDF documents*. In Proceedings of the IEEE International Conference on Bioinformatics and Biomedicine (BIBM). Atlanta, GA, USA. pp. 578-581.

Microsoft. Computer Vision. *Image Processing with the Computer Vision API*. https://azure.microsoft.com/en-us/services/cognitive-services/computer-vision/ (accessed March 11, 2019).

Milosevic, N., C. Gregson, R. Hernandez, and G. Nenadic. 2019. A framework for information extraction from tables in biomedical literature. *International Journal on Document Analysis and Recognition (IJDAR)* 22 (1): 55-78.

Nagy, G., and S. C. Seth. 1984. *Hierarchical representation of optically scanned documents*. In Proceedings of the 7th International Conference on Pattern Recognition (ICPR). Montreal, Canada. pp. 347-349.

Office Open XML. *Anatomy of a WordProcessingML File*. http://officeopenxml.com/anatomyofOOXML.php (accessed March 11, 2019).

Pick, J. L., S. Nakagawa, and D. W. A. Noble. 2019. Reproducible, flexible and high-throughput data extraction from primary literature: The MetaDigitise R package. *Methods in Ecology and Evolution* 10: 426-431.

Praczyk, P. A., and J. Nogueras-Iso. 2013. Automatic extraction of figures from scientific publications in high-energy physics. *Information Technology and Libraries* 32 (4): 25-52.

Rehman, A., and T. Saba. 2014. Neural networks for document image preprocessing: state of the art. *Artificial Intelligence Review* 42 (2): 253-273.

Savva, M., N. Kong, A. Chhajta, L. Fei-Fei, M. Agrawala, and J. Heer. 2011. *ReVision: automated classification, analysis and redesign of chart images*. In Proceedings of the 24th annual ACM symposium on User Interface Software and Technology (UIST). Santa Barbara, CA, USA. pp. 393-402.

Siegel, N., N. Lourie, R. Power, and W. Ammar. 2018. *Extracting Scientific Figures with Distantly Supervised Neural Networks*. In Proceedings of the 18th ACM/IEEE on Joint Conference on Digital Libraries (JCDL). Seatle, Washington,USA. pp. 223-232.

Siegel, N., Z. Horvitz, R. Levin, S. Divvala, and A. Farhadi. 2016. FigureSeer: Parsing result-figures in research papers. In *Computer Vision - ECCV 2016*. Lecture Notes in Computer Science. Ed. B. Leibe, J. Matas, N. Sebe, and M. Welling, pp. 664-680. Switzerland: Springer.

Smith, R. 2007. *An overview of the tesseract OCR engine*. In Proceedings of the Ninth International Conference on Document Analysis and Recognition. Parana, Brazil. pp. 629-633.

Srihari, S. N. 1986. *Document image understanding*. In Proceedings of the ACM Fall Joint Computer

Conference. Washington, D. C., USA. pp. 87-96.

Tsutsui, S., and D. J. Crandall. 2017. *A Data Driven Approach for Compound Figure Separation Using Convolutional Neural Networks*. In Proceedings of the 14th IAPR International Conference on Document Analysis and Recognition (ICDAR). Kyoto, Japan. pp. 533-540.

Williams, K., L. Li, M. Khabsa, J. Wu, P. C. Shih, and C. L. Giles. 2014. *A Web Service for Scholarly Big Data Information Extraction.* In Proceedings of the IEEE International Conference on Web Services. Anchorage, AK, USA. pp. 105-112.

Wu, J., K. M. Williams, H.-H. Chen, M. Khabsa, C. Caragea, S. Tuarob, A. G. Ororbia, D. Jordan, P. Mitra, and C. L. Giles. 2015. CiteSeerX: AI in a digital library search engine. *AI Magazine* 36 (3): 35-48.

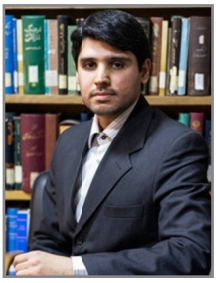
Xu, S., J. McCusker, and M. Krauthammer. 2008. Yale Image Finder (YIF): a new search engine for retrieving biomedical images. *Bioinformatics* 24 (17): 1968-1970.

Yang, X., E. Yumer, P. Asente, M. Kraley, D. Kifer, and C. L. Giles. 2017. *Learning to Extract Semantic Structure from Documents Using Multimodal Fully Convolutional Neural Networks*. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, HI, USA. pp. 4342-4351.

Yu, Y., H. Lin, J. Meng, X. Wei, and Z. Zhao. 2017. Assembling Deep Neural Networks for Medical Compound Figure Detection. *Information* 8 (2): 48-58.

آزاده فخرزاده

دارای مدرک تحصیلی دکتری در رشتهٔ پردازش تصویر از دانشگاه اوپسالای سوئد است. ایشان هم‌اکنون استادیار پژوهشکدهٔ فناوری اطلاعات، پژوهشگاه علوم و فناوری اطلاعات ایران (ایرانداک) است. پردازش تصویر، یادگیری ماشین، کلان‌داده‌ها، و یادگیری عمیق از جمله علایق پژوهشی وی است.



امیرحسین صدیقی

دانش‌آموختهٔ دکتری تخصصی در رشتهٔ مهندسی صنایع از دانشگاه صنعتی امیرکبیر (پلی‌تکنیک تهران) است. ایشان هم‌اکنون استادیار گروه پژوهشی سیستم‌های اطلاعاتی پژوهشگاه علوم و فناوری اطلاعات ایران است. بهینه‌سازی، سیستم‌های اطلاعاتی، کلان‌داده، شبکه‌های عصبی مصنوعی، و سیستم‌های انرژی از جمله علایق پژوهشی وی است.