Extracción de información de documentos PDF para su uso en la indización automática de *e-books*

Extracting information from PDF documents for use in automatic indexing of e-books

Isidoro GIL-LEIVA[1] 0000-0002-7175-3099

Mariângela Spotti Lopes FUJITA[2] 0000-0002-8239-7114

Franciele Marques REDIGOLO[3] 0000-0001-6277-2960

Jordan Ferreira SARAN[2] 0000-0002-8060-2268

Resumen

El número de libros electrónicos que ingresan en las bibliotecas en formato PDF cada día es mayor, complicando y haciendo casi inviables algunos procesos realizados tradicionalmente de forma manual por los bibliotecarios, como es la asignación de materias. En este contexto, se hace necesario el diseño y desarrollo de aplicaciones que asistan a los bibliotecarios. Teniendo esto en consideración, presentamos en este trabajo la evaluación de herramientas de extracción de información de libros en PDF que podrían usarse posteriormente como materia prima para un sistema de indización automática. Para ello, realizamos una primera evaluación de cinco softwares (PDFMiner.six, PDFAct, PDF-extract, PDFExtract y Grobib) y, posteriormente, como PDFAct consiguió el mejor rendimiento, hicimos una segunda evaluación para averiguar su capacidad para identificar y extraer informaciones de los libros, tales como títulos, índices, secciones, títulos de tablas y gráficos y referencias bibliográficas, informaciones relevantes para cualquier sistema de indización. Se concluye que ninguna de las herramientas evaluadas extrae adecuadamente las diferentes partes de libros en PDF, si bien, PDFAct ha logrado un rendimiento superior al del resto.

Palabras clave: Evaluación de software. Grobib. Indización automática. PDFMiner.six. PDFAct. PDF-extract. PDFExtract.

Abstract

The number of electronic books that enter libraries in PDF format is greater every day. Complicating and making it almost unfeasible for some processes, traditionally carried out manually by librarians such as the assignment of subjects, to be done. In this context, it is necessary to design and develop applications that assist librarians. Taking this into consideration, we present in this work the evaluation of tools for extracting information from books in PDF format that could be used later as raw material for an automatic indexing system. To do this, we carried out a first evaluation of five software (PDFMiner.six, PDFAct, PDF-extract, PDFExtract, and Grobib), later, as PDFAct achieved the best performance, we did a second evaluation to find out their ability to identify and extract information from the books such as titles, indexes, sections, titles of tables and graphs and bibliographic reference which are relevant information for any indexing

[1] Universidad de Murcia, Facultad de Comunicación y Documentación. *Campus* Universitario de Espinardo, s/n., 30100. Murcia, España. Correspondencia para/*Correspondence to*: I. GIL-LEIVA. *E-mail*: <isgil@um.es>.

[2] Universidade Estadual Paulista, Júlio de Mesquita Filho, Faculdade de Filosofia e Ciências, Programa de Pós-Graduação em Ciência da Informação. Marília, SP, Brasil.

[3] Universidade Federal do Pará, Faculdade de Biblioteconomia, Programa de Pós-Graduação em Ciência da Informação. Belém, PA, Brasil.

Recibido el 27 de octubre de 2021 e aprovado en 17 de novembro de 2021.

Cómo citar este articulo/*How to cite this article*

Gil-Leiva, I. *et al.* Extracción de información de documentos PDF para su uso en la indización automática de *e-books*. *Transinformação*, v. 34, e210069, 2022. https://doi.org/10.1590/2318-0889202234e210069

CC BY

system. It is concluded that none of the evaluated tools adequately extracts the different parts of PDF books, although PDFAct has achieved a better performance than the rest.

Keywords: *Software evaluation. PDFMiner.six. PDFAct. PDF-extract. PDFExtract. Grobib. Automatic indexing.*

Introducción

La cantidad de documentos disponibles en la actualidad es ingente, ya sea de documentos estructurados (tablas), semiestructurados (XML) o no estructurados (texto plano). Este volumen de documentos hace bastante ineficiente un procesamiento manual, debido al coste/persona y al tiempo. En las últimas décadas, se viene investigando en la Extracción de Información (EI) para fines diversos con el foco en la web, las bases de datos y en documentos PDF.

Desde principios de la década de 1990, una parte importante de los documentos electrónicos se almacenan y difunden en formato PDF y ya se cuentan por billones los documentos existentes en este formato. El PDF dificulta un procesamiento directo de la información contenida en él, de ahí la cantidad de investigaciones sobre EI de documentos PDF, tales como artículos científicos y libros.

Como la EI de documentos es una tarea compleja, es habitual dividir el proceso en diversas tareas, lo que permite seleccionar distintos procedimientos y algoritmos que mejor se adapten para su resolución. Cuando se pretende extraer información de artículos científicos o libros académicos, por ejemplo, dichos procedimientos y algoritmos se dividen para procesar, por una parte, la primera o primeras páginas de los documentos, de donde se extraen datos tales como títulos, nombres de personas, afiliación, direcciones, tablas de contenido/índices (en adelante, TOC), *etc.*; y, por otro lado, el procesamiento y extracción de información del resto del documento para identificar y extraer nombres de secciones, leyendas de tablas y gráficos, datos de las tablas, así como de las conclusiones o las referencias bibliográficas.

La estructura y el diseño de los documentos desempeña un rol importante en la EI. Así pues, un artículo científico, un *currículum vitae* e incluso una tesis doctoral, por lo general, tienen una estructura y diseño más homogéneo que los libros académicos. Un libro puede tener o no tener elementos como el TOC, un prólogo, una introducción, capítulos escritos por autores diferentes, referencias bibliográficas en cada capítulo, un único apartado de referencias al final del texto, lo que dificulta crear enfoques generales para la EI de este formato.

Para identificar y extraer información se usan diferentes elementos aportados por la estructura, el diseño o el propio texto. Así, la presencia de determinados elementos en el texto, tales como palabras con todas las letras en mayúsculas, palabras con la letra inicial en mayúscula, comillas (“”), guiones, signos de puntuación (,.:), números (19xx; 20xx; 2(4)), meses, determinadas palabras clave como ‘DOI’, ‘eds.’, ‘actas’, ‘pp.’ u otras señales como “se concluye”, “el método usado”, etc. son pistas que alertan de aspectos concretos.

Por otro lado, para la EI de documentos PDF se emplean principalmente dos enfoques diferentes: un enfoque basado en reglas que proceden de la observación como, por ejemplo, “el título de un libro antecede a los nombres de los autores”, “la afiliación de un autor va después que su nombre” o esta otra: “si en dos líneas de texto se encuentran dos de estos elementos ‘proceedings’, ‘pp.’, ‘p.’, ‘DOI’, ‘eds.’ se trata de una referencia bibliográfica”. Las aproximaciones basadas en reglas suelen tener un buen rendimiento porque derivan de la observación humana, pero son difíciles de adaptar a otros entornos diferentes en los que nacieron. Por otro lado, un segundo enfoque está basado en el aprendizaje automático, mediante la clasificación como el modelo oculto de Markov (HMM, por sus siglas en inglés), el campo aleatorio condicional, (CRF, por sus siglas en inglés) y la máquina de vectores de soporte (SVM, por sus siglas en inglés) con clasificadores de árbol de decisión, Bayes ingenuo o K vecinos más cercanos. Esta aproximación fundamentada en el aprendizaje automático, en general, tiene una adaptabilidad mayor que las propuestas de EI basadas en reglas.

La literatura sobre EI de documentos en PDF es extensa y variada. Parte de ella está dedicada a la EI de datos concretos de artículos científicos como, por ejemplo, información sobre citación y referencias bibliográficas (Dong *et al.*, 2017; Tkaczyk *et al.*, 2018; Haviana; Subroto, 2019; Alamoudi *et al.*, 2021); o la extracción de TOC (Perez-Arriaga, Estrada; Abad-Mota 2016; Najah-Imane; R'emi; Sira 2019; Shahid; Islam 2020). La EI de libros en PDF ha interesado a Khusro, Latif y Ullah (2015); Chaniago; Khodra (2017); Alamoudi *et al.* (2021); las tesis doctorales a Ojokoh, Adewale y Falaki (2009); los informes a Bui, Del Fiol y Jonnalagadda (2016); y los *currícula vitae* a Sandanayake *et al.* (2018); Chaudary *et al.* (2020); Anggakusuma, Mawardi y Lauro (2020); Pudasaini *et al.* (2021). Y, por último, también encontramos diversos trabajos en donde se efectúa una revisión bibliográfica sobre EI como los de Jayaram y Sangeeta (2017), Nasar, Jaffry y Malik (2018), Zaman, Mahdin y Hussain (2020), por mencionar solamente algunos recientes.

Como ha sido señalado, la generación de documentos digitales es enorme, así como su incorporación masiva a determinadas unidades documentales. Las bibliotecas, en los últimos años, han incrementado exponencialmente el número de e-books disponibles, acrecentado aún más durante 2020 y 2021, debido a la pandemia mundial de COVID. En Gil-Leiva *et al.* (2020) ofrecimos algunos datos de este incremento; también hablamos de la imposibilidad cada vez mayor de realizar la asignación de materias a los e-books por parte de los bibliotecarios debido a su incorporación masiva y dejando su asignación en manos de las editoriales; y también preguntamos a bibliotecarios de diferentes países sobre si los procesos técnicos que ejecutan actualmente sobre los libros electrónicos les producía cierto 'estrés' o 'frustración'. El 40% respondió que no, el 33% que no lo sabía y el 27% que sí. Por tanto, en sus respuestas se aprecia un rastro de estos sentimientos.

Por tanto, debido al imparable ingreso de e-books en las bibliotecas, se va haciendo necesario implementar sistemas que ayuden a los bibliotecarios a manejarse con decenas de este tipo de libros incorporados diariamente. Así pues, el trabajo que aquí se presenta se enmarca en este entorno descrito, donde la indización automática se podría convertir en una alidada importante para bibliotecarios y editores.

Desde finales de la década de 1950 se viene investigando intensamente en la automatización de la indización. Las metodologías empleadas han ido variando. En los primeros momentos, se utilizaba casi exclusivamente la estadística para obtener los términos de indización; pero, a partir de los años ochenta, se fueron incorporando técnicas de procesamiento del lenguaje natural. Sin embargo, lo habitual es que los prototipos incluyan una combinación de varias aproximaciones, como es el caso del cálculo de la frecuencia, reglas basadas en la posición que ocupan los términos en los documentos y herramientas, más o menos complejas, para el procesamiento del lenguaje natural (Gil-Leiva, 2008).

Objetivos

Considerando que la indización automática de documentos textuales usa como materia prima el texto, que los libros académicos son publicados principalmente en formato PDF, que es necesario usar herramientas para extraer el texto de los libros en PDF, este trabajo tiene como objetivo principal analizar herramientas de acceso libre y código abierto para la EI de libros en PDF que puedan usarse posteriormente en la indización automática de libros electrónicos. Al alcanzar este objetivo, daremos respuesta también a estas preguntas de investigación: ¿qué herramientas sirven para estos propósitos?, ¿qué herramientas permiten identificar y extraer de los libros académicos en PDF los títulos, TOC, leyendas de tablas y gráficos o las referencias bibliográficas, informaciones relevantes para una indización automática basada en reglas de posicionamiento de los términos?, ¿qué formatos de salida ofrecen las herramientas de EI de documentos PDF?

Procedimientos Metodologicos

En primer lugar, llevamos a cabo una búsqueda bibliográfica en la base de datos SCOPUS para incrementar nuestras lecturas y textos disponibles inicialmente. Realizamos una revisión de los resultados a través de los metadatos de los registros recuperados y procedimos a conseguir el texto completo de bastantes de ellos y a su análisis, con la intención de identificar herramientas/bibliotecas (en adelante, herramientas) de EI de documentos PDF. Esto nos permitió hallar finalmente cinco herramientas que sirvieron para realizar una primera evaluación de las mismas. Los datos logrados en esta primera evaluación nos permitieron seleccionar uno de ellos (PDFAct[4]) para efectuar una segunda evaluación más específica. En la Figura 1 se han resumido de forma esquemática estos procesos. Después de cada evaluación, tabulamos los resultados en tablas y efectuamos un análisis de los mismos y, finalmente, redactamos las conclusiones.

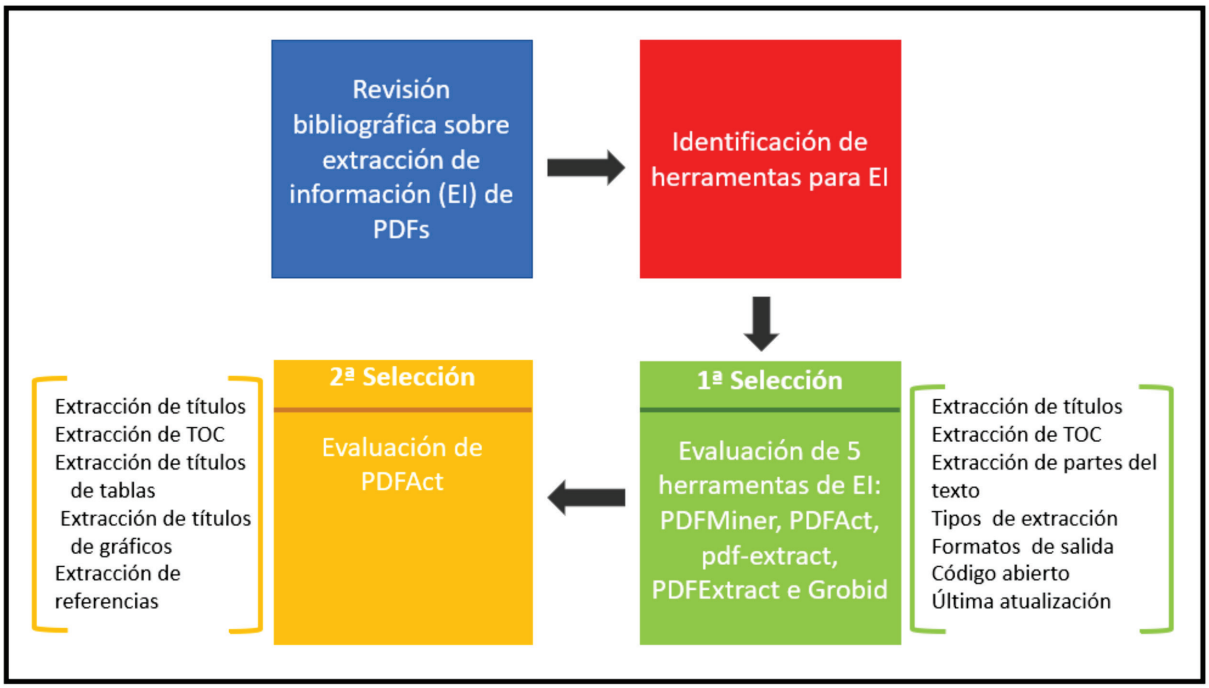

Figura 1 – Visión general de las fases.
Fuente: Elaboración de los autores (2021).

En la Base de datos Scopus usamos diferentes ecuaciones de búsqueda para localizar documentos sobre EI, tales como (TITLE-ABS-KEY ("information extraction") AND TITLE-ABS-KEY (e-books), (TITLE-ABS-KEY ("information extraction") AND TITLE-ABS-KEY (PDF)); (TITLE ("information extraction") AND ABS (e-books)), entre otras. Posteriormente, analizamos los registros recuperados para confirmar cuáles eran de nuestro interés. Así pues, localizamos numerosas propuestas y enfoques para la EI de documentos en PDF, así como diversos trabajos de revisión de EI. La lectura del texto completo de algunos de ellos nos permitió identificar en la parte de referencias bibliográficas otras investigaciones relevantes para nuestros objetivos. Un trabajo que resultó crucial para nosotros fue *Reconstructing scanned documents for full-text indexing to empower digital library services*, publicado en 2020 por Melania Nitu y sus colegas.

[4] PDFAct: https://github.com/ad-freiburg/PDFact

https://doi.org/10.1590/2318-0889202234e210069

En este trabajo se ofrece la descripción y características de once herramientas para la EI de documentos en PDF. De los once *softwares* presentados en Nitu *et al.* (2020) seleccionamos cinco herramientas que, por las especificaciones y funcionalidades señaladas, mejor se ajustaban a nuestro propósito principal de extraer información de e-books en PDF para su uso en la indización automática de los mismos; aspecto fundamental, puesto que el texto de salida de las herramientas de EI es el texto de entrada para los sistemas de indización automática. Con un texto de entrada de calidad se estará en mejores condiciones de lograr una indización automática de calidad, de ahí la importancia de acertar en la selección de la herramienta de EI. Las cinco herramientas seleccionadas son las que se mencionan a continuación.

Primera evaluación: PDFMiner.six[5], PDFAct, PDF-extract[6], PDFExtract[7] y Grobib[8].

Las variables que se analizaron para evaluar las cinco herramientas mencionadas son las siguientes: organización del texto de salida (estructurado, no estructurado); tipos de extracción del texto (bloques, palabras, párrafos, páginas o líneas); identificación de partes específicas del texto (títulos, TOCs, títulos de secciones, primer párrafo de cada sección, títulos de tablas y figuras y referencias bibliográficas; formatos de salida (XML, TXT, HTML, JSON, *etc.*); tipo de código fuente de la herramienta; en este último caso, si tiene código fuente abierto o cerrado; y, por último, la fecha de la última actualización de la herramienta.

Para llevar a cabo el experimento, creamos una colección de 20 libros en el campo de la informática. Todos los libros están en formato PDF y pertenecen a diferentes subáreas, tales como álgebra, programación, estadística y otras.

Para realizar el análisis se creó un entorno de desarrollo con el sistema operativo Big Sur (MacOS) y distribución Ubuntu. En cuanto a Big Sur OS, su configuración fue de 16GB de SRAM, (*Unified Memory*) procesador M1 (Apple Silicon) y 256 GB de SSD, y para Ubuntu OS, su configuración fue un procesador de 2 núcleos: 16 GB de RAM y 40 GB de SSSD. Se instaló y configuró según las instrucciones de los autores de cada herramienta.

A la hora de definir los parámetros para el análisis de las herramientas, conviene explicar con más detalle los procedimientos adoptados para tal fin. Realizamos un análisis intelectual y exploratorio de cada herramienta, donde configuramos cada entorno, ya fuera *Big Sur* OS o Ubuntu, según la documentación de cada herramienta. Los experimentos se realizaron con cada libro PDF por separado, realizando cada una de las pruebas y analizando si cada herramienta generaba resultados de acuerdo con los parámetros descritos. Se considera una extracción válida si lo extraído coincide con la información y partes de los libros en cuestión.

Para evaluar la extracción de TOC analizamos la documentación disponible buscando cualquier comando o parámetro disponible para realizar este tipo de acción. Si la herramienta lograba extraer TOC de todos los documentos, entonces se consideraba válida para esta tarea. Procedimos del mismo modo para tratar de extraer otras partes del texto como título, subtítulo y los elementos.

Para que una herramienta se considere apta con relación al tipo de extracción del texto, debe ser capaz de hacerlo en al menos dos formas (Ver Cuadro 1).

En cuanto al formato de salida tras la extracción del texto, al menos debería poder generar archivos en formato XML o JSON. Otra condición también analizada fue si las herramientas tenían su código fuente abierto o cerrado. Priorizamos las de código abierto y actualizaciones recientes porque permiten modificaciones dentro del propio código, si fuera necesario.

[5] Disponíble en: https://github.com/PDFminer/PDFminer.six
[6] Disponíble en: https://github.com/bitextor/PDF-extract
[7] Disponíble en: https://github.com/CrossRef/PDFextract
[8] Disponíble en: https://grobid.readthedocs.io/en/latest/

Para cuantificar la evaluación de las cinco herramientas definimos estas reglas:

- extrae el texto de forma estructurada, obtiene 1 punto;
- identifica el TOC dentro del texto: 1 punto;
- identifica el título, secciones, *etc*., 1 punto por cada parte extraída;
- dos tipos diferentes de extracción de texto: 1 punto;
- dos tipos de formatos de salida tras la EI: 1 punto;
- código fuente abierto: 1 punto;
- si hay empate, la herramienta con código fuente más reciente: 1 punto.

En las Tablas 3, 4 y 5 se muestran los resultados alcanzados. El software que obtuvo un mayor rendimiento fue seleccionado para ejecutar otra evaluación más específica. La herramienta seleccionada para esta segunda evaluación más detallada fue PDFAct. A continuación, se describe también la metodología seguida.

Segunda evaluación: PDFAct

Creamos una colección de veinte libros en PDF para cada uno de estos idiomas: inglés, portugués y español, y cada una de estas áreas: Ciencias de la computación, Ciencias de la información y Economía. Los elementos de estudio fueron: (a) título, (b) TOC, (c) títulos de tablas y figuras y (d) referencias.

Siguiendo la documentación de PDFAct, configuramos como tipo de extracción en "bloques", es decir, cada libro PDF consta de un bloque que contiene los siguientes metadatos:

- reglas del tipo (1) encabezado de texto; (2) título del libro; (3) TOC; (4) títulos de tablas y figuras; (5) cuerpo del texto; (6) referencias;
- posiciones, metadatos responsables de la información sobre los ejes X e Y de un bloque;
- página, que hace referencia a la página donde está presente el texto;
- texto, texto presente dentro del documento.

Seleccionamos también el formato JSON (*JavaScript Object Notation*) como una forma de conservar la información extraída del texto; sin embargo, la herramienta también proporciona otras opciones como HTML y TXT. Luego de configurar los tipos de metadatos para extraer el texto, realizamos un análisis intelectual en cada archivo JSON generado por PDFAct para verificar si era capaz de identificar cada uno de los aspectos a estudiar y convertimos esto en datos cuantitativos. Como puede observarse en la Tabla 4, las columnas contienen valores de 0 a 3. El valor 0 en la columna 'Título' significa que PDFAct no pudo identificar un título presente en un libro, el valor 1 se refiere a un porcentaje de similitud de hasta el 49% entre el título original y el título extraído. El valor 2 señala valores de similitud entre el 50% y el 74%, y el valor 3 apunta a una similitud del 75% o superior. Los porcentajes de similitud se generaron utilizando el algoritmo de Ratcliff y Metzener (1988) y una biblioteca llamada difflib[9] desarrollada en Python.

La columna TOC contiene dos tipos de valores: el valor 0 indica que PDFAct no pudo identificar ningún tipo de estructura próxima a un TOC; el valor 1 indica que en algún lugar del archivo JSON hay una estructura cerca de un TOC. Para obtener un valor de 0 o 1 en las columnas de título de tablas, figuras y referencias, se siguió el mismo procedimiento descrito para TOC.

Resultados

A continuación, presentamos los resultados obtenidos en las dos evaluaciones llevadas a cabo. En primer lugar, la evaluación de los cinco softwares de extracción de información de documentos PDFs: PDFMiner.six,

[9] https://docs.python.org/3/library/difflib.html

PDFAct, PDF-extract, PDFExtract y Grobib; y, seguidamente, los resultados de una segunda evaluación de PDFAct para comprobar si se adapta a nuestras necesidades volcadas en la indización automática de libros en PDF.

Primera evaluación: PDFMiner.six, PDFAct, PDF-extract, PDFExtract y Grobib

El Cuadro 1 muestra las partes analizadas con cada aplicación. Como se puede observar, PDF-extract, PDFAct y GROBID presentan un mejor desempeño en varios ítems.

Cuadro 1 – Análisis de las cinco herramientas.

Software	Extracción de TOC	Extracción de partes del texto	Tipos de extracción	Organización semántica	Formatos conversión	Código fuente abierto	Última actualización
PDFMiner.six	-	-	párrafos	-	XML, HTML, TXT	sí	Marzo/2021
PDFAct	-	X	palabras, bloques, líneas, páginas, párrafos	X	XML, TXT, JSON	sí	Mayo/2021
pdf-extract	-	X	párrafos	-	XML	sí	Septiembre/2015
PDFExtract	-	-	páginas, párrafos	X	HTML	sí	Abril/2021
GROBID	-	X	párrafos	X	XML	sí	Junio/2021

Fuente: Elaboración propia (2021).

En la Cuadro 2 se ofrece información sobre las partes de los documentos que cada software pudo extraer o en las que pudo realizar algún tipo de marca; si bien, en muchas ocasiones, sin mucha precisión.

Al observar el Cuadro 2, verificamos que ningún software pudo extraer el TOC, secciones o subsecciones, ni primeros párrafos.

Cuadro 2 – Extracción de partes del texto.

Partes de extracción del texto	PDFMiner.six	PDFAct	Pdf-extract	PDFExtract	GROBID
TOC	-	-	-	-	-
Título	-	X	X	-	X
Subtítulo	-	-	-	-	-
Secciones y subsecciones	-	X	-	-	-
Primer párrafo de una sección / subsección	-	-	-	-	-
Títulos de tablas y gráficos	-	X	-	-	-
Referencias	-	X	X	-	X

Fuente: Elaboración propia (2021).

GROBID extrajo los títulos de forma parcial y en cuanto a las referencias también fue parcial y desorganizada. Algunas palabras estaban juntas ("1Introduction" o "TheInteligence"). Algunos aspectos positivos a resaltar de GROBID son que cuenta con una documentación satisfactoria. GROBID puede ser usado como una aplicación JAR (Java Archive) o como un *WebService* y, por último, que presenta el código fuente más reciente.

Con relación a PDF-extract, hay que añadir un código fuente desactualizado que tal vez está detrás de algunos problemas que surgieron durante las pruebas; genera un archivo en formato XML con una secuencia ilógica de párrafos, lo que dificulta la identificación de partes específicas del texto, incluso para un humano.

PDFExtract, a pesar de no mostrar varios resultados satisfactorios, destaca por una mejor extracción del texto en general, en comparación con los textos originales. Esta herramienta podría ser eficaz para aplicar el TF-IDF y, en este sentido, parece diferenciarla del resto. Estas características señaladas se dan para los textos en inglés, español y portugués.

De acuerdo con lo Cuadro 2, PDFAct es la herramienta que obtuvo mejores resultados. Resultó igualmente la más eficaz para identificar partes del texto (Cuadro 3).

Cuadro 3 – Puntuación lograda por cada herramienta/biblioteca.

Herramienta/biblioteca	Puntuación
PDFMiner.six	3
PDFAct	9
Pdf-extract	5
PDFExtract	3
GROBID	5

Fuente: Elaboración propia (2021).

A pesar de que PDFAct logró el mejor rendimiento, no pudo identificar los TOC; sin embargo, el texto completo que obtiene es útil, contando con pocos casos de roturas aleatorias de líneas. Algunos aspectos negativos de PDFAct son los siguientes:

1) cuando los libros por capítulo contienen referencias, no puede identificar por separado cada conjunto de referencias;

2) devuelve texto contenido dentro de tablas y figuras;

3) consume mucha RAM para analizar cada libro;

4) a pesar de contar con una función específica para "TOC", no pudo extraer ninguno.

Por otro lado, PDFAct tiene de positivo que permite extraer texto de diferentes formas, tales como bloques, párrafos, líneas, páginas y palabras. En cualquier caso, de las herramientas analizadas, PDFAct parece mostrar las funcionalidades que mejor podrían adaptarse a nuestro proyecto de indización automática de libros en PDF. Los resultados obtenidos en la evaluación específica de PDFAct se presentan a continuación.

Evaluación de PDFAct

En el Cuadro 4 se presentan, de manera conjunta, los resultados obtenidos por PDFAct para la extracción de las diferentes partes de los documentos y para los tres idiomas. Al observar en la tabla las primeras cuatro columnas relativas al idioma portugués, constatamos que no ha podido identificar ningún TOC ni referencias. También es posible constatar que la herramienta logró identificar solo el 55% de los títulos y apenas un título de tabla o gráfico de un libro.

Los resultados de PDFAct para libros en inglés fueron más satisfactorios que para el portugués y el español. Las columnas I (idioma inglés), señalan que PDFAct fue el 100% eficaz en la identificación de títulos, resaltando que no hallamos un valor de 0 en ninguna de las columnas de Título. Los datos referidos a títulos de tablas y gráficos señalan una identificación del 80% y del 45% para referencias.

Con relación a la colección de libros en español (columnas E), se obtuvieron los peores resultados, si bien identificó el 50% de los títulos.

Cuadro 4 – Rendimiento de PDFAct para diferentes partes de los documentos por idiomas.

	P	P	P	P		I	I	I	I		E	E	E	E
Corpus	T	TOC	L	R		T	TOC	L	R		T	TOC	L	R
Libro 1	3	0	0	0		2	0	1	1		1	0	0	0
Libro 2	1	0	0	0		2	0	1	1		0	0	0	0
Libro 3	0	0	0	0		3	0	1	1		0	0	0	0
Libro 4	2	0	0	0		3	0	0	0		0	0	0	0
Libro 5	0	0	1	0		3	0	1	0		0	0	0	0
Libro 6	0	0	0	0		3	0	1	0		3	0	0	0
Libro 7	2	0	0	0		2	0	1	0		3	0	0	0
Libro 8	3	0	0	0		2	0	0	0		1	0	0	0
Libro 9	0	0	0	0		2	0	1	1		2	0	0	0
Libro 10	1	0	0	0		2	0	1	1		2	0	0	0
Libro 11	3	0	0	0		3	0	0	0		1	0	0	0
Libro 12	0	0	0	0		1	0	1	0		0	0	0	0
Libro 13	0	0	0	0		3	0	1	0		0	0	0	0
Libro 14	0	0	0	0		3	0	0	0		0	0	0	0
Libro 15	2	0	0	0		2	0	1	0		1	0	0	0
Libro 16	0	0	0	0		3	0	1	1		1	0	0	0
Libro 17	2	0	0	0		1	0	1	1		0	0	0	0
Libro 18	2	0	0	0		2	0	1	1		0	0	0	0
Libro 19	0	0	0	0		1	0	1	1		3	0	0	0
Libro 20	3	0	0	0		1	0	1	0		0	0	0	0
Total	11	0	1	0		20	0	16	9		10	0	0	0
Calcetín (%)	55	0	5	0		100	0	80	45		50	0	0	0

Note: P: idioma portugués; I: idioma inglés; E: idioma español; T: título; TOC: Table of Contents; L: leyendas de tablas y figuras; R: referencias.
Fuente: Elaboración propia (2021).

Tal vez lo más sobresaliente al examinar el Cuadro 4 es que PDFAct no es eficaz para extraer TOC y solamente parece válido para extraer el título, títulos de tablas y gráficos y referencias para el idioma inglés.

Conclusión y trabajos futuros

El incremento exponencial de la publicación de libros en PDF y la llegada masiva de e-books a las bibliotecas en los últimos años, hace que debamos investigar el diseño y desarrollo de herramientas que se puedan adaptar a los procesos bibliotecarios o de las editoriales para automatizar la asignación de descriptores o materias o, al menos, asistir en su ejecución. El trabajo presentado aquí ha analizado diversas herramientas para la EI con el fin de averiguar cuáles ofrecen las funcionalidades que mejor podrían adaptarse a un sistema de indización automática de libros electrónicos académicos. Los resultados logrados han permitido cumplir el objetivo marcado inicialmente, así como responder a las preguntas de investigación planteadas.

Se ha puesto de manifiesto que ninguna de las herramientas evaluadas logra una alta precisión en las tareas encomendadas. En cualquier caso, PDFAct ha logrado un rendimiento superior al del resto, si bien está lejos de ser una herramienta multilingüe completamente válida para identificar y extraer de libros en PDF títulos, TOC, títulos de tablas y gráficos o referencias bibliográficas, información de suma importancia para una indización automática fundamentada en reglas, a partir de la posición que ocupan los términos en los documentos.

En trabajos futuros, sería conveniente repetir los experimentos con una colección de documentos más grande; seguir buscando herramientas que ofrezcan un texto de salida totalmente reutilizable y, de seguir logrando

en otros ensayos datos similares a los presentados aquí, una posible alternativa podría ser mejorar el rendimiento de PDFAct con la implementación de varios algoritmos específicos que tuvieran como punto de partida la información identificada y extraída por esta herramienta.

Agradecimientos

Esta investigación ha sido financiada por *Fundação de Amparo à Pesquisa do Estado de São Paulo* en convenio de cooperación con FAPs/FAPESPA – *Fundação Amazônia de Amparo a Estudos e Pesquisas do Estado do Pará* (Proceso 2019/25470-6).

Contribuidores

I. GIL-LEIVA, M. S. L. FUJITA e F. M. REDIGOLO fueron responsables del diseño de la investigación y la redacción final; I. GIL-LEIVA fue el responsable de la revisión bibliográfica, sistematización y análisis de los resultados. J. F. SARAN fue responsable de la búsqueda bibliográfica y consolidación de los resultados. Todos los autores son responsables de la redacción final.

Referencias

Alamoudi, A. *et al.* A rule-based information extraction approach for extracting metadata from PDF books. *ICIC Express Letters, Part B: Applications*, v. 12, n. 2, p. 121-132, 2021. Doi: https://doi.org/ 10.24507/icicelb.12.02.121

Anggakusuma, J.; Mawardi, V.C.; Lauro, M.D. Resume extraction with conditional random field method. *IOP Conference Series*: *Materials Science and Engineering*, v. 1007, n. 1, 012154. 2020. Doi: https://doi.org/10.1088/1757-899X/1007/1/012154

Bui, D. D. A.; Del Fiol, G.; Jonnalagadda, S. PDF text classification to leverage information extraction from publication reports. *Journal of Biomedical Informatics*, v. 61, p. 141-148, 2016.

Chaniago, R.; Khodra, M. Information extraction on novel text using machine learning and rule-based system. *In*: International Conference on Innovative and Creative Information Technology, 2017. [*S.l.*]. *Proceedings* […]. [*S.l.*]: IEEE Explore, 2017. p. 1-6.

Chaudary, A. *et al.* Extraction of useful information from Crude Job Descriptions. *In*: IEEE International Multi-Topic Conference, INMIC, 23rd., 2020, Bahawalpur. *Proceedings* […]. [*S.l.*]: IEEE Explore, 2020. p. 1-4. Doi: https://doi.org/10.1109/INMIC50486.2020.9318132

Dong, A. *et al.* Citation Metadata Extraction via Deep Neural Network-based Segment Sequence Labeling. *In*: Conference on Information and Knowledge Management, 2017. Singapore. *Proceedings* […]. [*S./.*]: ACM, 2017. p. 1967-1970. Doi: https://doi.org/10.1145/3132847.3133074

Gil-Leiva, I. *Manual de indización*: teoría y práctica. Gijón: Trea, 2008.

Gil-Leiva, I. *et al.* The abandonment of the assignment of subject headings and classification codes in University Libraries due to the massive emergence of electronic books. *Knowledge Organization*, v. 47, n. 8, p. 646-667. 2020. Doi: https://doi.org/10.5771/0943-7444-2020-8-646

Haviana, S.; Subroto, I. Obtaining reference's topic congruity in Indonesian publications using machine learning approach. 2019. *In*: International Conference on Electrical Engineering, Computer Science and Informatics (EECSI), 6., 2019 [*S.l.*]. *Proceedings* […]. [*S.l.*:*s.n.*]: 2019. p. 428-431. Doi: https:// doi.org/10.23919/EECSI48112.2019.8976985

Jayaram, K.; Sangeeta, K. A review: Information extraction techniques from research papers. 2017. *In*: IEEE International Conference on Innovative Mechanisms for Industry Applications, 2017, Bengaluru, India. *Proceedings* […]. New York: IEEE, 2017. p. 56-59. Doi: https://doi.org/10.1109/ICIMIA.2017.7975532

Khusro, S.; Latif, A.; Ullah, I. On methods and tools of table detection, extraction and annotation in PDF documents. *Journal of Information Science*, v. 41, n. 1, p. 41-57, 2015. Doi: https://doi.org/10.1177/0165551514551903

Najah-Imane, B.; R'emi, J.; Sira, F. Table-of-contents generation on contemporary documents. *In*: International Conference on Document Analysis and Recognition (ICDAR), 15th., 2019, Sydney, Australia, september 20-25, 2019. *Proceedings* […]. New York: IEEE, 2019. p. 100-107. Doi: https://doi.org/10.1109/ICDAR.2019.00025

Nasar, Z.; Jaffry, S. W.; Malik, M. K. Information extraction from scientific articles: a survey. *Scientometrics*, v. 117, n. 3, p. 1931-1990, 2018. Doi: https://doi.org/10.1007/s11192-018-2921-5

Nitu, M. *et al.* Reconstructing scanned documents for full-text indexing to empower digital library services. *Lecture Notes in Computer Science* (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), v. 11984 LNCS, p. 183-190, 2020.

Ojokoh, B. A.; Adewale, O. S.; Falaki, S.O. Automated document metadata extraction. *Journal of Information Science*, v. 35, n. 5, p. 563-570, 2009. Doi: https://doi.org/10.1177/0165551509105195

Perez-Arriaga, M.O.; Estrada, T.; Abad-Mota, S. Tao: system for table detection and extraction from PDF documents. *In*: Markov, Z.; Russell, I. (ed.). *Proceedings of the Twenty-Ninth*

International Florida Artificial Intelligence Research Society Conference, *FLAIRS 2016, Key Largo*, Florida, May 16-18, 2016. Palo Alto: AAAI Press, 2016. p. 591-596.

Pudasaini, S. *et al.* Application of NLP for information extraction from unstructured documents. *Lecture Notes in Networks and Systems*, v. 209, p. 695-704, 2021. Doi: https://doi.org/10.1007/978-981-16-2126-0_54

Ratcliff, J. W.; Metzener, D. E. Pattern matching: the gestalt approach. *Dr. Dobb's Journal*, v. 13, n. 7, p. 46, 1988.

Sandanayake, T. C. *et al.* Automated CV analyzing and ranking tool to select candidates for job positions. *In*: Proceedings of the 6th International Conference on Information Technology: IoT and Smart City. 2018, Hong Kong. *Proceedings* […]. New York, NY: Association for Computing Machinery, 2018. p. 13-18. Doi: https://doi.org/10.1145/3301551.3301579

Shahid, M. H.; Islam, M. A. TOC generation in PDF Document for smart automated compliance engine. *In*: International Symposium on Recent Advances in Electrical Engineering & Computer Sciences (RAEE & CS), 2020, p. 1-5, Islamabad, Pakistan. *Proceedings* […]. New York: IEEE, 2020. Doi: https://doi.org/10.1109/raeecs50817.2020.9265792

Tkaczyk, D. *et al.* Machine learning vs. rules and out-of-the-box vs. retrained: an evaluation of open-source bibliographic reference and citation parsers. *In*: ACM/IEEE on Joint Conference on Digital Libraries, 18., June 3-7, 2018, Fort Worth, Texas, USA. *Proceedings* […]. New York, NY: Association for Computing Machinery, 2018. https://doi.org/10.1145/3197026.3197048

Zaman, G.; Mahdin, H.; Hussain, K. Information extraction from semi and unstructured data sources: a systematic literature review. *ICIC Express Letters*, v. 14, n. 6, p. 593-603, 2020. Doi: https://doi.org/10.24507/icicel.14.06.593