Extracción de modelos de conocimiento a partir de libros de texto y su aplicación en los negocios

Isaac Alpizar-Chacon
Escuela de Administración
de Tecnologías de Información
Instituto Tecnológico de Costa Rica, Costa Rica
ialpizar@tec.ac.cr

Resumen

Este artículo describe un enfoque innovador para la extracción de modelos de conocimiento a partir de libros de texto y su aplicación en los negocios. A través de un proceso detallado, que incluye la extracción, vinculación, enriquecimiento, análisis y formalización, se crean modelos de conocimiento de alta calidad. Este proceso es el resultado de una investigación doctoral, y en este artículo se ofrece un resumen ejecutivo de la misma. Los modelos creados tienen aplicaciones en la educación, pero también pueden ser usados potencialmente en el ámbito empresarial. En ese sentido, este artículo explora también las posibilidades de adaptar este proceso para interpretar y procesar documentos empresariales esenciales.

Palabras clave: libros de texto, vinculación de información, análisis de conocimiento, documentos empresariales

Introducción

En la era digital actual, muchos sistemas de información, como los sistemas educativos o de negocio, giran en torno al conocimiento. Los sistemas educativos adaptativos e inteligentes, por ejemplo, necesitan representaciones de conocimiento de alta calidad para funcionar eficazmente. Pero, ¿de dónde obtenemos este conocimiento y cómo lo presentamos de una manera que las máquinas puedan entenderlo?

Adquirir conocimiento en el formato correcto se considera el proceso más largo y complejo en el desarrollo de sistemas que requieren conocimiento [1]. Normalmente, el conocimiento se obtiene de expertos humanos utilizando diversas técnicas [2].

Sin embargo, el proceso es muy laborioso y consume mucho tiempo. Además, los expertos suelen tener dificultades para articular su conocimiento y tienden a estar en desacuerdo.

Los documentos de texto se utilizan típicamente como fuente para la extracción automática de conocimiento [3, 4]. Sin embargo, a pesar de la amplia disponibilidad de recursos textuales en todos los dominios de conocimiento, los métodos escalables de extracción de conocimiento para crear modelos para un dominio específico aún son un problema abierto.

Una vía para resolver el problema es la utilización de los libros de texto digitales, ya que son una fuente increíblemente rica de conocimiento. Los libros de texto están llenos de información sobre una amplia gama de temas, presentada de una manera estructurada y fácil de seguir. Sin embargo, extraer este conocimiento de los libros de texto (en PDF) y convertirlo en un formato que las máquinas puedan utilizar es un desafío considerable.

Este artículo hace un resumen de una tesis doctoral [5] enfocada en abordar el problema descrito. El enfoque presentado extrae automáticamente modelos de conocimiento a partir de libros de texto digitales. Los modelos creados pueden ser leídos automáticamente por las máquinas. Además, este artículo también presenta una discusión de cómo se podrían utilizar los modelos de conocimiento generados en el área de los negocios.

Libros de texto

Los libros de texto son ricos en elementos estructurales y de contenido que facilitan el aprendizaje. Su material principal, diseñado para transmitir información a los estudiantes, está organizado coherente y jerárquicamente, reflejando tanto el dominio disciplinario como el conocimiento del estudiante [6, 7]. La Tabla de Contenidos (T.C.) actúa como una guía de navegación, mostrando la disposición jerárquica de capítulos y subcapítulos, y proporcionando una visión general de los temas tratados [8]. Un ejemplo de T.C. se ilustra en la Figura 1. Por otro lado, el índice, situado al final del libro, lista y ordena términos esenciales, con entradas que a menudo tienen una estructura jerárquica y referencias cruzadas [9, 10]. La Figura 2 muestra un ejemplo de índice. Además de estos elementos, los libros de texto incluyen encabezados, números de página, estilos de formato, y otros componentes como imágenes que enriquecen la experiencia de lectura y aprendizaje.

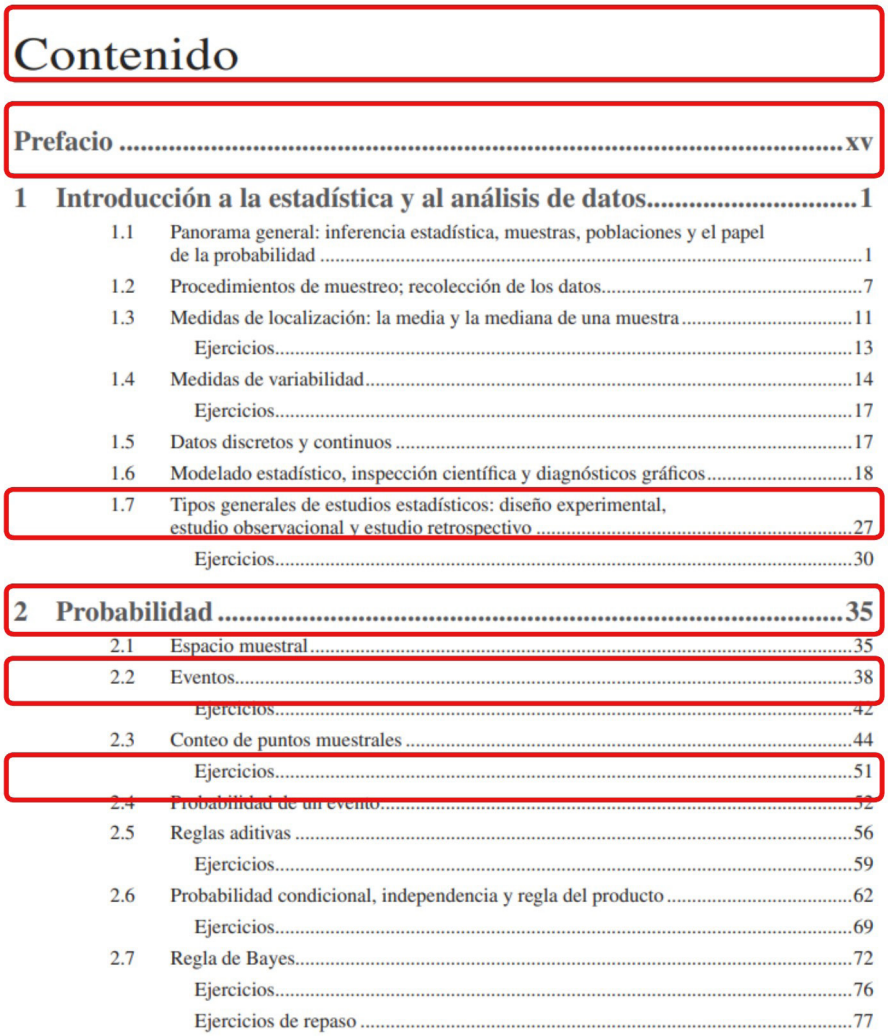

Figura 1. Ejemplo de los elementos identificados en una Tabla de Contenidos (T.C.).

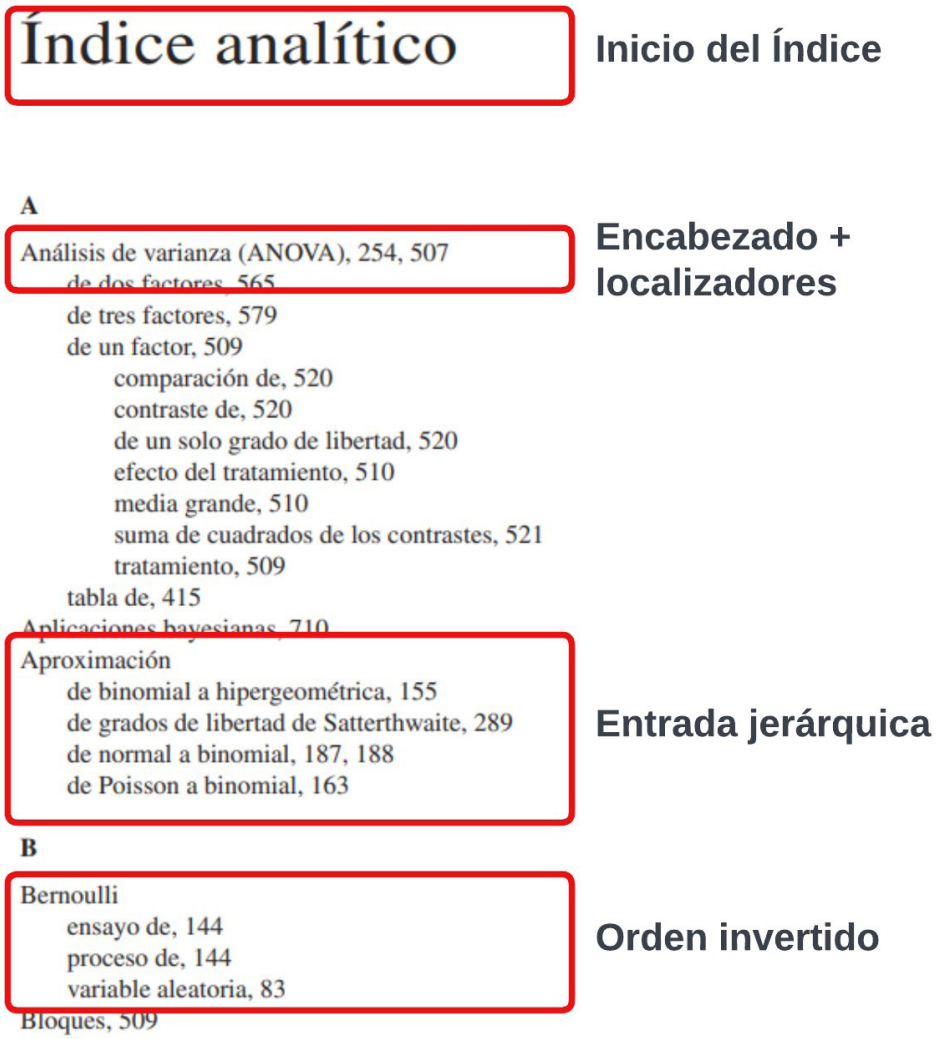

Figura 2. Ejemplo de los elementos identificados en el índice.

Proceso

Los autores aplican su expertise al redactar los libros de texto, con el objetivo de simplificar y explicar la materia a los estudiantes. Al examinar los elementos de formato y estructurales en textos de alta calidad, es posible decodificar automáticamente la información en ellos contenida y, gradualmente, adquirir entendimiento sobre un área específica. Esta sección presenta la metodología propuesta para extraer, vincular, enriquecer, analizar y formalizar modelos de conocimiento a partir de libros en formato PDF. Estos modelos son representaciones de alta calidad de un área específica de conocimiento (un dominio).

La Figura 3 presenta el proceso para la extracción de modelos de conocimiento a partir de libros de texto. El proceso tiene varias entradas, salidas, fases, etapas y pasos.

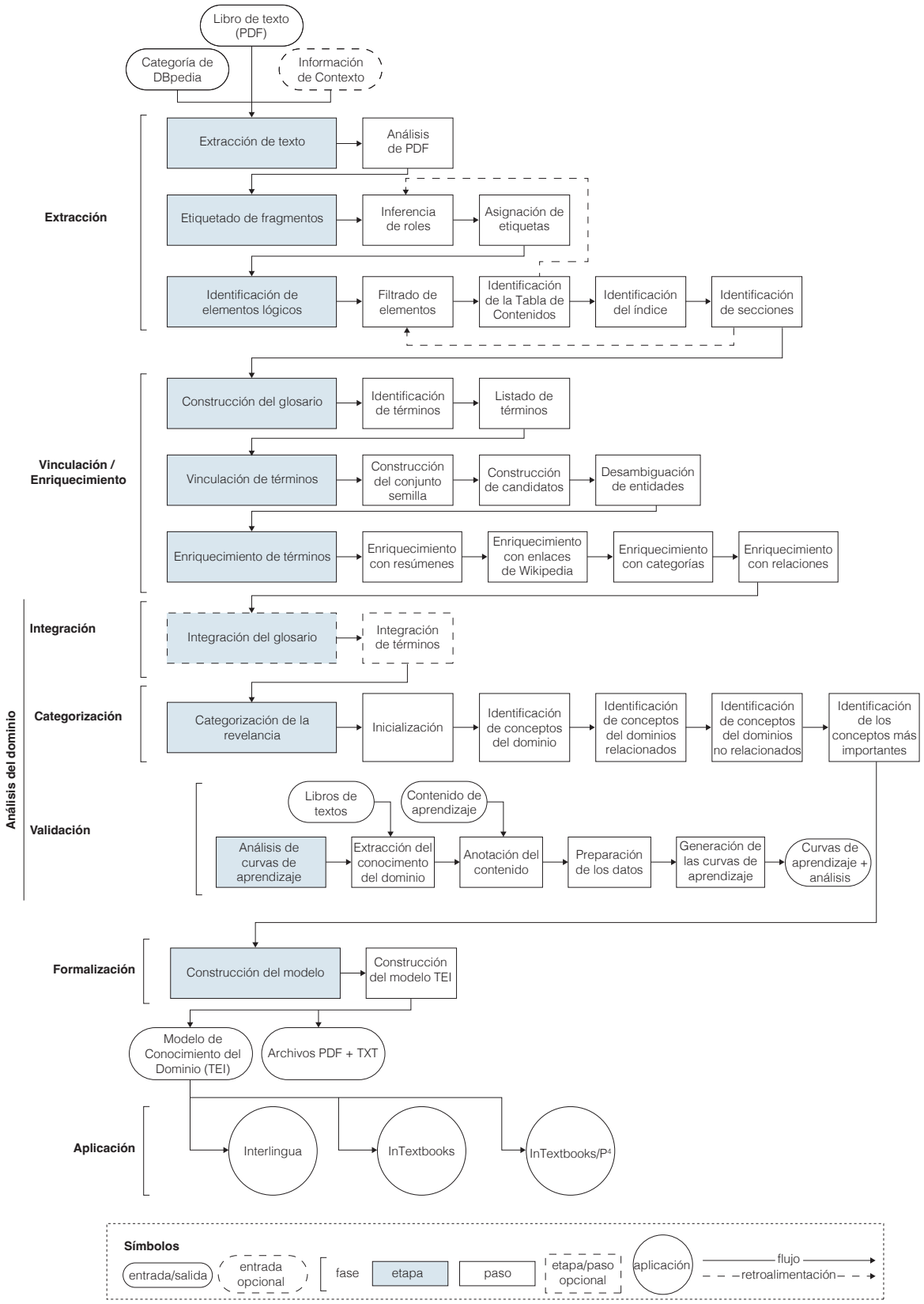

Figura 3. Proceso para la extracción de modelos de conocimiento a partir de libros de texto.

En la primera fase del enfoque (extracción), se extrae la estructura, el contenido y los términos del dominio de un libro de texto. La información estructural contiene la lista de capítulos y subcapítulos del libro de texto. El contenido del libro de texto se representa de manera estructurada (palabras, líneas, párrafos, páginas y secciones). Por último, los términos del dominio se extraen del índice al final del libro, que contiene la terminología utilizada en el libro de texto y el dominio.

En la siguiente fase (vinculación / enriquecimiento), los términos del dominio se utilizan como un puente para vincular los libros de texto a una base de conocimiento externa. Específicamente, los términos del dominio se relacionan con entidades en DBpedia[1], la cual es una base de conocimientos basada en Wikipedia.

En la tercera, cuarta y quinta fase se realiza un análisis del dominio. Primero (integración), los términos de varios libros de texto se integran en un solo modelo para obtener una mejor cobertura del dominio. Luego (categorización), los términos se categorizan según su relevancia para el dominio para identificar los conceptos que pertenecen al dominio principal del libro de texto, a dominios relacionados o a dominios no relacionados. Después de eso (validación), se establece la validez de los conceptos extraídos como elementos para representar y evaluar el conocimiento en el dominio.

La fase final del enfoque automático es la formalización, donde todo el conocimiento extraído se serializa como un archivo XML (utilizando el formato *Text Encoding Initiative*[2]).

Una vez que el enfoque ha producido los modelos de conocimiento, estos están listos para ser utilizados en varias aplicaciones. En pocas palabras, el método sugerido comienza extrayendo datos básicos de los libros, como estructura, contenido y términos. Después, se añade más información poco a poco, como enlaces y contenido con significado específico. Finalmente, se analiza y se mejora el entendimiento sobre el tema, enfocándose en conceptos clave. El resultado son modelos que nos ayudan a entender mejor un tema específico.

Resultados

Los modelos de conocimiento producidos son una representación de conocimiento de alta calidad. Estos modelos poseen seis propiedades:

1. Precisión: La información del libro de texto se representa adecuadamente en los modelos de conocimiento. En general, el contenido, los elementos de la Tabla de Contenidos y los términos en el índice se extraen de forma precisa [11, 12].
2. Semántica: Los modelos de conocimiento incorporan información adicional extraída de una base de conocimiento externa. Por lo general, los términos del índice se vinculan o conectan eficazmente con los elementos de DBpedia, y por lo tanto, indirectamente con Wikipedia [12, 13].
3. Cobertura: Los modelos de conocimiento abarcan una porción significativa del dominio. En general, cuando se utilizan varios libros de texto, se logra una representación extensa del área de conocimiento de interés [12, 13].
4. Relevancia: Los conceptos en los modelos de conocimiento poseen una relevancia identificada para el dominio. Por lo general, se pueden distinguir los conceptos que pertenecen al área de conocimiento de interés, aquellos que pertenecen a áreas de conocimiento relacionadas y los que corresponden a áreas no relacionadas [14].
5. Validez cognitiva: Los conceptos en los modelos de conocimiento pueden utilizarse para modelar y evaluar el conocimiento en el área de interés [15].

1 https://www.dbpedia.org/

2 https://tei-c.org/

6. Granularidad: Los conceptos en los modelos de conocimiento son componentes concretos que pueden utilizarse para modelar y evaluar el conocimiento en el área de interés [15].

Los modelos de conocimiento generados facilitan la conexión entre diversos contenidos, eliminando intervenciones manuales y permitiendo vinculaciones entre secciones y glosarios similares [16]. Son esenciales en Sistemas Educativos Adaptativos e Inteligentes, proporcionando adaptaciones personalizadas y conexiones lingüísticas entre idiomas [17, 18]. Además, su aplicabilidad se extiende a áreas como el mundo empresarial, donde se necesita conocimiento de dominio de alta calidad.

Aplicación en los negocios

El proceso propuesto puede conceptualizarse en dos aplicaciones fundamentales dentro del ámbito empresarial. La primera aplicación se dedica a crear modelos de conocimiento específicos para el dominio de los negocios, utilizando libros de texto relevantes, como *Introducción a los Negocios* [19]. Estos modelos tienen aplicaciones tanto educativas, para facilitar la enseñanza, como industriales, donde pueden ser integrados en sistemas avanzados de información y herramientas de Inteligencia Artificial para soluciones como *chatbots* o análisis predictivo.

La segunda aplicación aborda el procesamiento de documentos esenciales en negocios, tales como informes financieros y contratos. Dada la importancia de los datos en texto libre en el ámbito comercial [20], se han desarrollado métodos de procesamiento de texto para extraer conceptos de documentos comerciales [21] y visualizar información de patentes [22]. Una fortaleza del proceso propuesto es su sistema de reglas adaptable, lo que permite una alta precisión en el reconocimiento de documentos empresariales específicos.

En conjunto, la adaptación del proceso en el ámbito empresarial promete innovaciones en la gestión de información y educación en el sector. Los modelos de conocimiento, derivados de fuentes de calidad, pueden impulsar eficiencia, servicios personalizados y decisiones basadas en datos, brindando una ventaja en el mercado actual.

Conclusión y trabajo futuro

La extracción de modelos de conocimiento a partir de libros de texto representa una oportunidad significativa en la era digital. El proceso propuesto en este documento demuestra ser versátil y preciso, con aplicaciones potenciales en diversos dominios. La aplicación en el ámbito empresarial, en particular, abre nuevas vías para la interpretación y procesamiento de documentos clave, lo que puede conducir a una mayor eficiencia y efectividad en las operaciones comerciales. La metodología presentada aquí brinda una base sólida para futuras investigaciones y desarrollos en este campo.

Referencias

[1] Pham, D.T., y Dimov, S.S. (1997). An efficient algorithm for automatic knowledge acquisition. *Pattern Recognition*, 30(7), 1137–1143.

[2] Schreiber, G. (2008). Knowledge engineering. *Foundations of Artificial Intelligence*, 3, 929–946.

[3] Rau, L.F., Jacobs, P.S., Zernik, U. (1989). Information extraction and text summarization using linguistic knowledge acquisition. *Information Processing & Management,* 25(4), 419–428.

[4] Ferreira-Mello, R., André, M., Pinheiro, A., Costa, E., Romero, C. (2019). Text mining in education. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *9*(6), e1332.

[5] Alpizar Chacon, I. (2023). *Extraction of knowledge models from textbooks.* (Thesis Doctoral, Utrecht University). https://doi.org/10.33540/1647

[6] Chiappetta, E.L., Fillman, D.A., Sethna, G.H. (1991). A method to quantify major themes of scientific literacy in science textbooks. *Journal of research in science teaching*, *28*(8), 713–725.

[7] Chambliss, M.J., y Calfee, R.C. (1989). Designing science textbooks to enhance student understanding. *Educational psychologist*, 24(3), 307–322.

[8] Déjean, H., y Meunier, J.-L. (2009). On tables of contents and how to recognize them. *International Journal of Document Analysis and Recognition (IJDAR),* 12(1), 1–20.

[9] *The chicago manual of style* (17.a ed.). (2017). The University of Chicago Press.

[10] Asaishi, T. (2011). An analysis of the terminological structure of index terms in textbooks. *Procedia-Social and Behavioral Sciences*, 27, 209–217.

[11] Alpizar-Chacon, I., y Sosnovsky, S. (2020). Order out of chaos: Construction of knowledge models from pdf textbooks. *Proceedings of the acm symposium on document engineering 2020* (pp. 1–10).

[12] Alpizar-Chacon, I., y Sosnovsky, S. (2021). Knowledge models from pdf textbooks. *New Review of Hypermedia and Multimedia*, 27(1-2), 128–176.

[13] Alpizar-Chacon, I., y Sosnovsky, S. (2019a). Expanding the web of knowledge: one textbook at a time. *Proceedings of the 30th on hypertext and social media*. ACM.

[14] Alpizar-Chacon, I., y Sosnovsky, S. (2022). What's in an index: Extracting domainspecific knowledge graphs from textbooks. *Proceedings of the acm web conference 2022 (www '22)* (pp. 966–976).

[15] Alpizar-Chacon, I., Sosnovsky, S., Brusilovsky, P. (2023). Measuring the quality of domain models extracted from textbooks with learning curves analysis. *International conference on artificial intelligence in education* (pp. 804–809).

[16] Alpizar-Chacon, I., Barria-Pineda, J., Akhuseyinoglu, K., Sosnovsky, S., Brusilovsky, P. (2021). Integrating textbooks with smart interactive content for learning programming. *Proceedings of the third workshop on intelligent textbooks* (Vol. 2895, pp. 4–18). CEUR WS.

[17] Alpizar-Chacon, I., y Sosnovsky, S. (2019b). Interlingua: Linking textbooks across different languages. *Proceedings of the first workshop on intelligent textbooks* (Vol. 2384, p. 104-117). CEUR-WS.

[18] Alpizar-Chacon, I., van der Hart, M., Wiersma, Z.S., Theunissen, L.S., Sosnovsky, S. (2020). Transformation of pdf textbooks into intelligent educational resources. *Proceedings of the second workshop on intelligent textbooks* (Vol. 2674, pp. 4–16). CEUR-WS.

[19] Pride, W.M., Hughes, R.J., Kapoor, J.R., Aurora, Z.E.A. (2017*). Introducción a los negocios*. Cengage Learning Editores.

[20] Abramowicz, W., y Piskorski, J. (2003). Information extraction from free-text business documents. *Effective databases for text & document management* (pp. 12–23). IGI Global.

[21] Ménard, P.A., y Ratté, S. (2016). Concept extraction from business documents for software engineering projects. *Automated Software Engineering*, 23(4), 649–686.

[22] Dražić, M., Kukolj, D., Vitas, M., Pokrić, M., Manojlović, S., Tekić, Z. (2013). Effectiveness of text processing in patent documents visualization. *2013 ieee 11th international symposium on intelligent systems and informatics (sisy)* (pp. 287–291).

Sobre el autor

Isaac Alpízar-Chacón

El Dr. Isaac Alpízar Chacón es profesor adjunto a tiempo parcial en la Escuela de Administración de Tecnologías de Información. Posee un doctorado en Ciencias de la Información y la Computación de la Universidad de Utrecht, Países Bajos, así como una Maestría en Ciencias de la Computación de la Universidad del Sarre, en Saarbrücken, Alemania. Sus áreas de interés son la investigación educativa y la inteligencia artificial en la educación. ORCID: https://orcid.org/0000-0002-6931-9787.