7 REFERENCES

[1] H. Chao and J. Fan, “Layout and Content Extraction for PDF Documents,” *document analysis systems,* pp. 213-224, 2004.

[2] K. Hadjar, M. Rigamonti, D. Lalanne and R. Ingold, “Xed: a new tool for extracting hidden structures from electronic documents,” in *First International Workshop on Document Image Analysis for Libraries, 2004. Proceedings.*, 2004.

[3] E. Wustner, T. Hotzel and P. Buxmann, “Converting business documents:a classification of problems and solutions using XML/XSLT,” in *Proceedings Fourth IEEE International Workshop on Advanced Issues of E-Commerce and Web-Based Information Systems (WECWIS 2002),* 2002.

[4] A. Dengel and F. Dubiel, “Clustering and classification of document structure-a machine learning approach,” in *Proceedings of 3rd International Conference on Document Analysis and Recognition,* 1995.

[5] A. Anjewierden and S. Kabel, “Automatic indexing of PDF documents with ontologies,” , 2001.

[6] A. Anjewierden, “AIDAS: incremental logical structure discovery in PDF documents,” in *Proceedings of Sixth International Conference on Document Analysis and Recognition,* 2001.

[7] H. L. Chieu, H. T. Ng and Y. K. Lee, “Closing the Gap: Learning-Based Information Extraction Rivaling Knowledge-Engineering Methods,” in *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics,* 2003.

[8] D. Esser, D. Schuster, K. Muthmann, M. Berger and A. Schill, “Automatic indexing of scanned documents: a layout-based approach,” in *Document Recognition and Retrieval XIX*, 2012.

[9] C. Ramakrishnan, A. Patnia, E. H. Hovy and G. A. P. C. Burns, “Layout-aware text extraction from full-text PDF of scientific articles,” *Source Code for Biology and Medicine*, vol. 7, no. 1, pp. 7-7, 2012.

[10] R. Futrelle, M. Shao, C. Cieslik and A. Grimes, “Extraction,layout analysis and classification of diagrams in PDF documents,” in *Seventh International Conference on Document Analysis and Recognition, 2003. Proceedings.*, 2003.

[11] R. Futrelle, “Ambiguity in visual language theory and its role in diagram parsing,” in *Proceedings 1999 IEEE Symposium on Visual Languages*, 1999.

[12] R. Futrelle and N. Nikolakis, “Efficient analysis of complex diagrams using constraint-based parsing,” in *Proceedings of 3rd International Conference on Document Analysis and Recognition*, 1995.

[13] S.-H. Lin and J.-M. Ho, “Discovering informative content blocks from Web documents,” in *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, 2002.

[14] H. Bast and C. Korzen, “A benchmark and evaluation for text extraction from PDF,” in *Proceedings of the 17th ACM/IEEE Joint Conference on Digital Libraries*, 2017.

[15] W. Kehong, “Optimized hierarchy clustering based extraction for logical document structures,” *Journal of Tsinghua University*, 2005.

[16] P. N. Smith and D. F. Brailsford, “Towards structured, block-based PDF,” , 1995.

[17] T. Padova, Adobe Acrobat 7 PDF Bible, 2001.

[18] R. Cohn, Portable Document Format Reference Manual, 1993.

[19] T. Joachims, Learning to Classify Text Using Support Vector Machines, 2002.

[20] A. Jain and B. Yu, “Document representation and its application to page decomposition,” *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 20, no. 3, pp. 294-308, 1998.

[21] T. Hu and R. Ingold, “A mixed approach toward an efficient logical structure recognition from document images,” *Electronic Publishing,* vol. 6, pp. 457-468, 1993.

[22] J. Fan, “Text extraction via an edge-bounded averaging and a parametric character model,” in *Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series*, 2003.

[23] M. Lipinski, K. Yao, C. Breitinger, J. Beel and B. Gipp, “Evaluation of header metadata extraction approaches and tools for scientific PDF documents,” in *Proceedings of the 13th ACM/IEEE-CS joint conference on Digital libraries,* 2013.

[24] Y. Wang, I. T. Phillips and R. M. Haralick, “A Study on the Document Zone Content Classification Problem,” *document analysis systems,* pp. 212-223, 2002.

[25] L. Zhang, Y. Pan and T. Zhang, “Focused named entity recognition using machine learning,” in *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, 2004.

[26] E. Saund, “Scientific challenges underlying production document processing,” in *Proceedings of SPIE, the International Society for Optical Engineering,* 2011.

[27] R. Futrelle, “Strategies for diagram understanding: generalized equivalence, spatial/object pyramids and animate vision,” in *[1990] Proceedings. 10th International Conference on Pattern Recognition,* 1990.

[28] E. A. El-Kwae and K. H. Atmakuri, “Document image representation using XML technologies,” in *Document Recognition and Retrieval IX*, 2001.

[29] H. Déjean and J.-L. Meunier, "A system for converting PDF documents into structured XML format," *document analysis systems,* pp. 129-140, 2006.

[30] D. Tkaczyk, A. Czeczko, K. Rusek, L. Bolikowski and R. Bogacewicz, "GROTOAP: ground truth for open access publications," in *Proceedings of the 12th ACM/IEEE-CS joint conference on Digital Libraries*, 2012.