

BÖLÜM 9

BİLGİ ÇIKARIMI İÇİN BELGE DÜZENİ ÇÖZÜMLEME YÖNTEMLERİ

Emine KISKANÇ¹
Hasan H. APIK²
Meltem KALAYCI³
Vedat AYBAR⁴
Faruk SERİN⁵

GİRİŞ

İnsanlar var olduğu günden bu yana içinde buldukları dönemin teknolojiyle yaşadıklarını kayıt altına almışlardır. Bilim ve teknolojinin gelişmesi ve mürekkep, kalem, kâğıt gibi yazı nesnelerinin icat edilmesiyle çeşitli belgelerin oluşturulması kolaylaşmış ve yaygınlaşmıştır. Devletlerin kurumsallaşması ile vatandaşların kurumlarda yaptıkları işlemleri hızlandırmak ve belirli bir standarda oturtmak için doküman/belgelerin kullanılması bir zorunluluk haline gelmiştir.

Bilgisayarların icat edilmesine kadar geçen sürede belgeler, kurumların arşivlerinde yazılı biçimde korunurdu. Bu arşivler fiziksel ortam olduğundan ve fazla kopyalar oluşturmak maliyetli olduğu için belgelerin korunması güçleşebilmekteydi. Matbaanın icat edilmesiyle kopya sayılarını artırmak nispeten kolaylaşmıştır. Bilgisayar ve internetin icat edilmesiyle belgelerin dijital kopyalarının oluşturulması ve bu kopyaların saklanması çok daha az maliyetlerle sağlanmaktadır.

Günümüzde dijital ortamda belgeler oluşturulabildiği gibi çok eski belgeler taratılarak dijital ortama aktarılabilir. Dijital ortamda oluşturulan belgeler algoritmalar tarafından daha kolay anlaşılabilir fakat yazılı belgeler bilgisayara aktarıldığında algoritmaların bunları çözümleyebilmesi için bazı ön işlemler gerekmektedir. Bu nedenle hem dijital ortamda oluşturulmuş belgelerin hem de basılı ortamdan dijital ortama aktarılan belgelerin analiz edilmesi başlı başına

¹ Bilgisayar Görmesi Mühendisi, Mobildev, emine.kiskanc@mobildev.com

² Bilgi Teknolojileri Direktörü, Mobildev, hasan.apik@mobildev.com

³ Ürün Yöneticisi, Mobildev, meltem.kalayci@mobildev.com

⁴ Pazarlama Direktörü, Mobildev, vedat.aybar@mobildev.com

⁵ Doç. Dr., Mersin Üniversitesi, Mühendislik Fakültesi, Bilgisayar Mühendisliği, fserin@mersin.edu.tr

fatura, makbuz, başvuru formu, tablo, resim, metin verisi tanıma gibi çözümler sunmaktadır.

SONUÇ

Kullanılan belge sayısı arttıkça bu belgelerden bilgiyi otomatik ve doğru bir şekilde çıkarabilmek, gittikçe daha kritik bir konuya dönüşmektedir. Özellikle belgeler homojen bir yapıya sahip değil ve içerisinde düzenli olmayan bir formatta başka alt belgeler bulunuyorsa, bu alt belgelerin ayrıştırılması gerekebilmektedir. Bu durum, literatürde genel olarak belge düzeni çözümleme olarak ele alınmaktadır. Bu konuyla ilgili bazı problemler çözülmüş ve farklı amaçlar için akademik çalışmalar yapılmış, açık kaynak kodlar sunulmuş ve ticari ürünler geliştirilmiştir. Bununla beraber yaklaşımların ve ürünlerin genelleştirilebilmesi ve kullanımlarının yaygınlaştırılabilmesi için geliştirme ihtiyaçları devam etmektedir. Bu çalışmada, bu konudaki çalışmalara katkıda bulunabilecek bilgiler sunulmuştur.

KAYNAKLAR

1. G. M. Binmakhshen ve S. A. Mahmoud, "Document Layout Analysis: A Comprehensive Survey", *ACM Comput. Surv.*, c. 52, sy 6, s. 109:1-109:36, Eki. 2019, doi: 10.1145/3355610.
2. K. Y. Wong, R. G. Casey, ve F. M. Wahl, "Document Analysis System", *IBM J. Res. Dev.*, c. 26, sy 6, ss. 647-656, Kas. 1982, doi: 10.1147/rd.266.0647.
3. M. Atay, M. Kalayci, H. Apik, V. Aybar, F. Serin, ve A. O. Akyüz, "An Approach to Analyzing the Layout of Unstructured Digital Documents", *30th Signal Processing and Communications Applications Conference (SIU)*, IEEE, 2022.
4. L. O'Gorman, "The document spectrum for page layout analysis", *IEEE Trans. Pattern Anal. Mach. Intell.*, c. 15, sy 11, ss. 1162-1173, Kas. 1993, doi: 10.1109/34.244677.
5. Jaekyu Ha, R. M. Haralick, ve I. T. Phillips, "Recursive X-Y cut using bounding boxes of connected components", *Proc. 3rd Int. Conf. Doc. Anal. Recognit.*, c. 2, ss. 952-955, 1995, doi: 10.1109/ICDAR.1995.602059.
6. S. Marinai, M. Gori, ve G. Soda, "Artificial neural networks for document analysis and recognition", *IEEE Trans. Pattern Anal. Mach. Intell.*, c. 27, sy 1, ss. 23-35, Oca. 2005, doi: 10.1109/TPAMI.2005.4.
7. M. Shilman, P. Liang, ve P. Viola, "Learning Non-Generative Grammatical Models for Document Analysis.", Oca. 2005, c. 2, ss. 962-969. doi: 10.1109/ICCV.2005.140.
8. H. Wei, M. Baechler, F. Slimane, ve R. Ingold, "Evaluation of SVM, MLP and GMM Classifiers for Layout Analysis of Historical Documents", içinde *2013 12th International Conference on Document Analysis and Recognition*, Ağu. 2013, ss. 1220-1224. doi: 10.1109/ICDAR.2013.247.
9. Y. Xu, M. Li, L. Cui, S. Huang, F. Wei, ve M. Zhou, "LayoutLM: Pre-training of Text and Layout for Document Image Understanding", içinde *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, Ağu. 2020, ss. 1192-1200. doi: 10.1145/3394486.3403172.
10. D. A. Borges Oliveira ve M. P. Viana, "Fast CNN-Based Document Layout Analysis", içinde *2017 IEEE International Conference on Computer Vision Workshops (ICCVW)*, Eki. 2017, ss. 1173-1180. doi: 10.1109/ICCVW.2017.142.
11. A. R. Katti vd., "Chargrid: Towards Understanding 2D Documents", içinde *Proceedings of the*

2018 *Conference on Empirical Methods in Natural Language Processing*, Brussels, Belgium, Eki 2018, ss. 4459-4469. doi: 10.18653/v1/D18-1476.

12. P. Forczmański, A. Smolinski, A. Nowosielski, ve K. Małecki, “Segmentation of Scanned Documents Using Deep-Learning Approach”, içinde *Advances in Intelligent Systems and Computing*, 2020, ss. 141-152. doi: 10.1007/978-3-030-19738-4_15.
13. P. Zhang *vd.*, “VSR: A Unified Framework for Document Layout Analysis combining Vision, Semantics and Relations”. arXiv, 13 Mayıs 2021. doi: 10.48550/arXiv.2105.06220.
14. X. Wu, Z. Hu, X. Du, J. Yang, ve L. He, “Document Layout Analysis via Dynamic Residual Feature Fusion”, içinde *2021 IEEE International Conference on Multimedia and Expo (ICME)*, Tem. 2021, ss. 1-6. doi: 10.1109/ICME51207.2021.9428465.
15. Y. Xu *vd.*, “LayoutLMv2: Multi-modal Pre-training for Visually-Rich Document Understanding”. arXiv, 09 Ocak 2022. doi: 10.48550/arXiv.2012.14740.
16. Y. Huang, T. Lv, L. Cui, Y. Lu, ve F. Wei, “LayoutLMv3: Pre-training for Document AI with Unified Text and Image Masking”. arXiv, 19 Temmuz 2022. doi: 10.48550/arXiv.2204.08387.
17. S. Li, X. Ma, S. Pan, J. Hu, L. Shi, ve Q. Wang, “VTLayOut: Fusion of Visual and Text Features for Document Layout Analysis”. arXiv, 12 Ağustos 2021. doi: 10.48550/arXiv.2108.13297.
18. Z. Shen, R. Zhang, M. Dell, B. C. G. Lee, J. Carlson, ve W. Li, “LayoutParser: A Unified Toolkit for Deep Learning Based Document Image Analysis”. arXiv, 21 Haziran 2021. doi: 10.48550/arXiv.2103.15348.
19. S. Long, S. Qin, D. Pantelev, A. Bissacco, Y. Fujii, ve M. Raptis, “Towards End-to-End Unified Scene Text Detection and Layout Analysis”, program adı: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, ss. 1049-1059. Erişim: 23 Aralık 2022. (Çevrimiçi). Erişim adresi: https://openaccess.thecvf.com/content/CVPR2022/html/Long_Towards_End-to-End_Unified_Scene_Text_Detection_and_Layout_Analysis_CVPR_2022_paper.html
20. S. Biswas, A. Banerjee, J. Lladós, ve U. Pal, “DocSegTr: An Instance-Level End-to-End Document Image Segmentation Transformer”. arXiv, 21 Eylül 2022. doi: 10.48550/arXiv.2201.11438.
21. J. Li, Y. Xu, T. Lv, L. Cui, C. Zhang, ve F. Wei, “DiT: Self-supervised Pre-training for Document Image Transformer”. arXiv, 19 Temmuz 2022. doi: 10.48550/arXiv.2203.02378.
22. R. Girshick, J. Donahue, T. Darrell, ve J. Malik, “Rich feature hierarchies for accurate object detection and semantic segmentation”. arXiv, 22 Ekim 2014. Erişim: 27 Aralık 2022. (Çevrimiçi). Erişim adresi: <http://arxiv.org/abs/1311.2524>
23. R. Girshick, “Fast R-CNN”. arXiv, 27 Eylül 2015. Erişim: 27 Aralık 2022. (Çevrimiçi). Erişim adresi: <http://arxiv.org/abs/1504.08083>
24. S. Ren, K. He, R. Girshick, ve J. Sun, “Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks”. arXiv, 06 Ocak 2016. Erişim: 27 Aralık 2022. (Çevrimiçi). Erişim adresi: <http://arxiv.org/abs/1506.01497>
25. J. Redmon, S. Divvala, R. Girshick, ve A. Farhadi, “You Only Look Once: Unified, Real-Time Object Detection”. arXiv, 09 Mayıs 2016. Erişim: 27 Aralık 2022. (Çevrimiçi). Erişim adresi: <http://arxiv.org/abs/1506.02640>
26. J. Redmon ve A. Farhadi, “YOLO9000: Better, Faster, Stronger”. arXiv, 25 Aralık 2016. Erişim: 27 Aralık 2022. (Çevrimiçi). Erişim adresi: <http://arxiv.org/abs/1612.08242>
27. J. Redmon ve A. Farhadi, “YOLOv3: An Incremental Improvement”. arXiv, 08 Nisan 2018. Erişim: 27 Aralık 2022. (Çevrimiçi). Erişim adresi: <http://arxiv.org/abs/1804.02767>
28. A. Bochkovskiy, C.-Y. Wang, ve H.-Y. M. Liao, “YOLOv4: Optimal Speed and Accuracy of Object Detection”. arXiv, 22 Nisan 2020. Erişim: 27 Aralık 2022. (Çevrimiçi). Erişim adresi: <http://arxiv.org/abs/2004.10934>
29. “YOLOv5 Documentation”. <https://docs.ultralytics.com/> (erişim 27 Aralık 2022).
30. “YOLOv5”. Ultralytics. (Çevrimiçi). Erişim adresi: <https://github.com/ultralytics/yolov5>
31. C. Li *vd.*, “YOLOv6: A Single-Stage Object Detection Framework for Industrial Applications”. arXiv, 07 Eylül 2022. Erişim: 27 Aralık 2022. (Çevrimiçi). Erişim adresi: <http://arxiv.org/>

- abs/2209.02976
32. C.-Y. Wang, A. Bochkovskiy, ve H.-Y. M. Liao, “YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors”. arXiv, 06 Temmuz 2022. Erişim: 27 Aralık 2022. (Çevrimiçi). Erişim adresi: <http://arxiv.org/abs/2207.02696>
 33. X. Zhong, J. Tang, ve A. J. Yepes, “PubLayNet: largest dataset ever for document layout analysis”. arXiv, 15 Ağustos 2019. doi: 10.48550/arXiv.1908.07836.
 34. B. Pfitzmann, C. Auer, M. Dolfi, A. S. Nassar, ve P. W. J. Staar, “DocLayNet: A Large Human-Annotated Dataset for Document-Layout Analysis”, içinde *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, Ağu. 2022, ss. 3743-3751. doi: 10.1145/3534678.3539043.
 35. A. Jimeno-Yepes, P. Zhong, ve D. Burdick, “ICDAR 2021 Competition on Scientific Literature Parsing”, 2021, ss. 605-617. doi: 10.1007/978-3-030-86337-1_40.
 36. A. Abdallah, A. Berendeyev, I. Nuradin, ve D. Nurseitov, “TNCR: Table Net Detection and Classification Dataset”, *Neurocomputing*, c. 473, ss. 79-97, Şub. 2022, doi: 10.1016/j.neucom.2021.11.101.
 37. H. Desai, P. Kayal, ve M. Singh, “TabLeX: A Benchmark Dataset for Structure and Content Information Extraction from Scientific Tables”, 2021, ss. 554-569. doi: 10.1007/978-3-030-86331-9_36.
 38. B. Smock, R. Pesala, ve R. Abraham, “PubTables-1M: Towards comprehensive table extraction from unstructured documents”. arXiv, 18 Kasım 2021. doi: 10.48550/arXiv.2110.00061.
 39. Z. Wang, Y. Xu, L. Cui, J. Shang, ve F. Wei, “LayoutReader: Pre-training of Text and Layout for Reading Order Detection”. arXiv, 27 Ağustos 2021. Erişim: 23 Aralık 2022. (Çevrimiçi). Erişim adresi: <http://arxiv.org/abs/2108.11591>
 40. M. Li, L. Cui, S. Huang, F. Wei, M. Zhou, ve Z. Li, “TableBank: Table Benchmark for Image-based Table Detection and Recognition”, içinde *Proceedings of the Twelfth Language Resources and Evaluation Conference*, Marseille, France, May. 2020, ss. 1918-1925. Erişim: 23 Aralık 2022. (Çevrimiçi). Erişim adresi: <https://aclanthology.org/2020.lrec-1.236>
 41. M. Li *vd.*, “DocBank: A Benchmark Dataset for Document Layout Analysis”, Oca. 2020, ss. 949-960. doi: 10.18653/v1/2020.coling-main.82.
 42. A. Mondal, P. Lipps, ve C. Jawahar, “IIIT-AR-13K: A New Dataset for Graphical Object Detection in Documents”, Ağu. 2020.
 43. L. Gao *vd.*, “ICDAR 2019 Competition on Table Detection and Recognition (cTDaR)”, 2019 *Int. Conf. Doc. Anal. Recognit. ICDAR*, ss. 1510-1515, 2019, doi: 10.1109/ICDAR.2019.00243.
 44. L. Cui, Y. Xu, T. Lv, ve F. Wei, “Document AI: Benchmarks, Models and Applications”. arXiv, 16 Kasım 2021. Erişim: 22 Aralık 2022. (Çevrimiçi). Erişim adresi: <http://arxiv.org/abs/2111.08609>
 45. “Intelligently Extract Text & Data with OCR - Amazon Textract - Amazon Web Services”, *Amazon Web Services, Inc.* <https://aws.amazon.com/textract/> (erişim 22 Aralık 2022).
 46. “Document AI”, *Google Cloud.* <https://cloud.google.com/document-ai> (erişim 22 Aralık 2022).
 47. “Form Recognizer – Automated Data Processing Systems | Microsoft Azure”. <https://azure.microsoft.com/en-us/products/form-recognizer/> (erişim 22 Aralık 2022).