III SHO‘BA. KOMPYUTER LINGVODIDAKTIKASI

AKADEMIK HUJJATLARNING FORMATI VA TUZILISHI

Turg‘unboyev Rashid
Qo‘qon davlat universiteti

Annotatsiya: Ushbu maqolada akademik hujjatlardan metama’lumotlarni avtomatik ekstraksiya qilish tizimlarini ishlab chiqishda hujjat formatlari va ularning ichki tuzilishini tushunishning ahamiyati tahlil qilinadi. PDF, LaTeX, XML (xususan JATS XML), HTML va EPUB kabi formatlarning afzalliklari, cheklovlari hamda metama’lumotlarni ajratishdagi o‘ziga xos muammolari ko‘rib chiqiladi. PDF formatining keng tarqalganligi va unda semantik strukturaning yo‘qligi, shuningdek, tug‘ma va skanerlangan PDF o‘rtasidagi farqlar alohida ta’kidlanadi. LaTeX formatining mantiqiy strukturani aniq belgilashdagi qulayliklari, XML formatining standartlashtirilganligi, ammo ularning cheklangan qo‘llanilishi haqida ma’lumot beriladi. Shuningdek, nashriyotlar shablonlaridagi xilma-xillik, formatlararo konversiya muammolari hamda ma’lumotlar to‘plamlarining (DocBank, TableBank) roli yoritiladi. Maqola metama’lumotlarni ajratish tizimlariga qo‘yiladigan asosiy talablarni aniqlash bilan yakunlanadi.

Kalit so‘zlar: *Akademik hujjatlar, metama’lumotlarni avtomatik ekstraksiya qilish, PDF, LaTeX, XML, JATS, HTML, hujjat tuzilishi, formatlararo konversiya, OCR, mashinaviy o‘rganish, DocBank, TableBank, nashriyot shablonlari, semantik tahlil*

Akademik maqolalardan metama’lumotlarni avtomatik ekstraksiya qilish tizimlarini ishlab chiqishda eng muhim omillardan biri bu hujjatlarning formatlari va ularning ichki tuzilish xususiyatlarini chuqur tushunishdir. Ilmiy nashrlar turli formatlarda mavjud bo‘lib, ularning har biri o‘ziga xos afzalliklar, cheklovlar va ma’lumotlarni ifodalash usullariga ega. Ushbu formatlarning xususiyatlarini bilmasdan turib, metama’lumotlarni samarali ekstraksiya qilishga qodir tizimni

yaratish deyarli imkonsizdir. Zamonaviy ilmiy axborot almashinuvida eng keng tarqalgan formatlar sifatida PDF, LaTeX, XML va uning ilmiy nashrlar uchun ixtisoslashtirilgan variantlari, shuningdek, HTML va EPUB kabi formatlarni ko‘rsatish mumkin.[1]

PDF formati bugungi kunda akademik nashrlar sohasida eng keng tarqalgan va eng ko‘p qo‘llaniladigan format hisoblanadi. Adobe kompaniyasi tomonidan ishlab chiqilgan ushbu format 1990-yillarning boshlaridan boshlab hujjatlarni platformadan mustaqil ravishda ko‘rsatish va chop etish uchun standartga aylangan. PDF formatining asosiy afzalligi shundaki, u hujjatning vizual ko‘rinishini - shriftlar, rasmlar, jadvallar, sahifalarning joylashuvi va boshqa dizayn elementlarini - muallif yoki nashriyot tomonidan belgilangan shaklda saqlab qoladi. Bu xususiyat PDFni nashriyotlar uchun juda qulay formatga aylantiradi, chunki maqola har qanday qurilmada va operatsion tizimda bir xil ko‘rinishda aks etadi. Biroq, aynan shu xususiyat metama’lumotlarni avtomatik ekstraksiya qilish nuqtai nazaridan jiddiy muammolarni keltirib chiqaradi.[2] PDF formatidagi hujjatlarda matn va boshqa elementlar asosan vizual joylashuv asosida ifodalanadi, ya’ni ularning mantiqiy strukturasi (bu qism sarlavha, bu qism muallif ismi, bu qism annotatsiya) haqida hech qanday ma’lumot mavjud emas. Dasturiy vositalar maqolaning sahifalaridagi matn qatorlari va bloklarining faqat geometrik koordinatalarini ko‘radi, ammo ularning semantik ma’nosini bilmaydi.

PDF formatining yana bir muhim xususiyati uning ikki xil turda mavjudligidir: “tug‘ma” PDF va “skanerlangan” PDF. Tug‘ma PDF hujjatlar dastlab elektron shaklda yaratilgan bo‘lib, ulardagi matnlar odatda to‘g‘ridan-to‘g‘ri ajratib olish mumkin. Bunga misol qilib, Microsoft Word yoki LaTeX dasturlarida tayyorlangan va PDF formatiga eksport qilingan maqolalarni keltirish mumkin. Skanerlangan PDF hujjatlar esa qog‘oz nusxalarni skanerlash orqali yaratiladi va ular aslida rasm ko‘rinishida bo‘ladi. Bunday hujjatlardan matnni ajratib olish uchun

OCR texnologiyasidan foydalanish talab etiladi.[3] Skanerlangan hujjatlar bilan ishlash qo‘shimcha murakkabliklarni keltirib chiqaradi, chunki OCR jarayonida matnni aniqlashda xatoliklar yuz berishi mumkin, shuningdek, asl hujjatdagi formatlash (qalin shrift, kursiv, shrift o‘lchamlari) haqidagi ma’lumotlar yo‘qolishi yoki noto‘g‘ri talqin qilinishi mumkin. Choudhury va boshqalar tomonidan olib borilgan tadqiqotda skanerlangan elektron tezis va dissertatsiyalardan metama’lumotlarni ajratish muammosi batafsil o‘rganilgan bo‘lib, mualliflar OCR natijalariga asoslangan evristik usullarni taklif qilganlar.[4]

LaTeX formati ilmiy-texnik hujjatlarni tayyorlash uchun maxsus ishlab chiqilgan matnli formatdir. U, ayniqsa, matematika, fizika, axborot texnologiyalari va muhandislik sohalarida keng qo‘llaniladi. LaTeX ning asosiy afzalligi shundaki, unda hujjatning mantiqiy strukturasi maxsus buyruqlar (masalan, \title, \author, \section, \begin{abstract}) yordamida aniq belgilanadi.[5] Bu xususiyat LaTeX hujjatlarini metama’lumotlarni avtomatik ajratish uchun juda qulay formatga aylantiradi, chunki sarlavha, muallif, annotatsiya va boshqa elementlarni ekstraksiya qilish uchun murakkab algoritmlar talab etilmaydi – tegishli teglarni qidirish kifoya. LaTeX ning yana bir muhim afzalligi shundaki, u matnli formatda bo‘lgani uchun har qanday matn muharririda ochilishi va qayta ishlanishi mumkin. Biroq, LaTeX hujjatlari ham o‘ziga xos kamchiliklarga ega. Birinchidan, ular asosan aniq fanlar sohasidagi maqolalar uchun qo‘llaniladi va barcha nashriyotlar tomonidan qabul qilinmaydi. Ikkinchidan, LaTeX hujjatlarida foydalanuvchi tomonidan belgilangan maxsus buyruqlar mavjud bo‘lishi mumkin, bu esa ularni qayta ishlashni murakkablashtiradi. Ling va Chen tomonidan ishlab chiqilgan DeepPaperComposer tizimi LaTeX hujjatlaridagi strukturaviy ma’lumotlardan foydalanib, konvolyutsion neyron tarmoqlarni o‘qitish uchun yuqori sifatli o‘quv ma’lumotlarini avtomatik yaratish imkonini beradi.[6] PDF, LaTeX va XML formatlarida bir xil maqola elementining (masalan, sarlavha) turlicha ifodalanishi 1-rasmda ko‘rsatilgan.

XML formati metama’lumotlarni ifodalashning eng tuzilgan va standartlashtirilgan usuli hisoblanadi. XML hujjatlarida ma’lumotlar maxsus teglar bilan belgilanadi, bu esa ma’lumotlarning semantik ma’nosini aniq ifodalash imkonini beradi. Ilmiy nashrlar sohasida eng keng tarqalgan XML standarti JATS hisoblanadi. JATS XML milliy tibbiyot kutubxonasi tomonidan ishlab chiqilgan bo‘lib, u ilmiy maqolalarning barcha tarkibiy elementlarini - sarlavha, mualliflar, annotatsiya, bo‘limlar, jadvallar, rasmlar, formulalar, adabiyotlar – standartlashtirilgan teglar bilan belgilash imkonini beradi.[7] JATS XML formatining afzalligi shundaki, u ma’lumotlarni mashinalar tomonidan o‘qilishi va qayta ishlanishini maksimal darajada osonlashtiradi. Ko‘plab yirik nashriyotlar (masalan, Elsevier, Springer, Wiley) va ilmiy ma’lumotlar bazalari (PubMed Central) o‘z maqolalarini JATS XML formatida saqlaydi va tarqatadi. Biroq, barcha nashriyotlar va jurnallar o‘z maqolalarini XML formatida taqdim etavermaydi. Ko‘pgina kichik nashriyotlar, konferensiya materiallari va institutsional repozitoriylar asosan PDF formatidan foydalanadi. Shu sababli, metama’lumotlarni ajratish tizimlari asosan PDF formatidagi maqolalar bilan ishlashga majbur.

PDF (visual)	LaTeX (mantiqiy)	XML (strukturaviy)
Akademik maqolalardan metama’lumotlarni ekstraksiya qilishning avtomatlashtirilgan tizimi	\title{Akademik maqolalardan metama’lumotlarni ekstraksiya qilishning avtomatlashtirilgan tizimi}	<article-meta> <title-group> <article-title> Akademik maqolalardan metama’lumotlarni ekstraksiya qilishning avtomatlashtirilgan tizimi </article-title> </title-group> </article-meta>

1-rasm. PDF, LaTeX va XML formatlarida maqola sarlavhasining strukturaviy ifodalanishi

HTML formati, ayniqsa, onlayn ilmiy jurnallar va repozitoriylarda keng qo‘llaniladi. HTML hujjatlari ham XML kabi teglar asosida qurilgan bo‘lib, ular veb-brauzerlarda ko‘rish uchun mo‘ljallangan. Ilmiy maqolalarning HTML versiyalari ko‘pincha nashriyot veb-saytlarida mavjud bo‘lib, ularda metama’lumotlar meta teglar yordamida alohida ajratilgan bo‘lishi mumkin. Biroq, HTML formatining standartlashtirish darajasi XML ga qaraganda pastroq, chunki turli nashriyotlar o‘zlarining xususiy dizayn va struktura elementlaridan foydalanadi. Shuningdek, HTML hujjatlari odatda maqolaning to‘liq matnini emas, balki uning veb-sahifada ko‘rinishini ifodalaydi.

Akademik hujjatlarning formatlari bilan bog‘liq yana bir muhim jihat - bu ularning strukturaviy murakkabligi va xilma-xilliligidir. Har bir nashriyot, hatto har bir jurnal o‘zining xususiy shablonlari va formatlash usullariga ega. Ba’zi jurnallar sarlavhani sahifaning yuqori qismida, katta shrift bilan, markazda joylashtirsa, boshqalari chap tomonda, kichikroq shrift bilan joylashtirishi mumkin. Muallif ismlari ba’zi jurnallarda “Mualliflar” sarlavhasi ostida, ba’zilarida esa to‘g‘ridan-to‘g‘ri sarlavhadan keyin keltiriladi. Affiliatsiya ma’lumotlari ba’zan muallif ismlari bilan bir qatorda, ba’zan esa alohida blokda, sahifaning pastki qismida joylashtiriladi. Annotatsiya ba’zi jurnallarda “Abstract” sarlavhasi bilan ajratilgan bo‘lsa, boshqalarida maxsus sarlavhasiz, faqat shrift o‘lchami yoki formatlash orqali farqlanadi. Bu xilma-xillik metama’lumotlarni avtomatik ekstraksiya qilish tizimlari uchun jiddiy muammo hisoblanadi. Boukhers va boshqalar tomonidan olib borilgan tadqiqotda nemis ijtimoiy fanlaridagi nashrlar misolida ushbu xilma-xillikning qanchalik yuqori ekanligi va an’anaviy NLP usullarining bunday hujjatlardan metama’lumotlarni aniq ajratib olishda yetarli samaradorlikka erisha olmasligi ko‘rsatilgan.[8] Mualliflar ushbu muammoni yengish uchun hujjatni rasm sifatida ko‘radigan va Mask R-CNN kabi kompyuter ko‘rish modellaridan foydalanadigan yondashuvni taklif qilganlar.

Turli formatlardagi akademik hujjatlardan metama’lumotlarni ekstraksiya qilishda yana bir muhim muammo – bu formatlar o‘rtasidagi konversiya jarayonidagi ma’lumot yo‘qotishlari va transformatsiya xatolaridir. Masalan, LaTeX dan PDF ga konversiya qilishda strukturaviy teglar yo‘qoladi va hujjat vizual ko‘rinishga aylanadi. PDF dan XML ga konversiya qilishda esa aksincha, vizual joylashuv asosida mantiqiy strukturani qayta tiklashga harakat qilinadi, bu esa har doim ham to‘liq va aniq bo‘lavermaydi. Iwashokun va Ade-Ibijola tomonidan ishlab chiqilgan RX parser kontekstga asoslangan grammatikalar va muntazam ifodalar yordamida tadqiqot hujjatlarini XML formatiga avtomatik ravishda tahlil qilish imkonini beradi.[7] Ushbu tizim 160 sahifagacha bo‘lgan turli xil hujjatlarda 91% muvaffaqiyat ko‘rsatkichiga erishgan.

Akademik hujjatlarning formatlari va tuzilishi masalasida yana bir muhim jihat – bu hujjatlarning ichki tarkibiy qismlari (bo‘limlar, jadvallar, rasmlar, formulalar) va ular orasidagi munosabatlardir. Ilmiy maqolalar odatda kirish, usullar, natijalar, muhokama va xulosalar kabi standart bo‘limlardan iborat. Biroq, bu bo‘limlarning nomlanishi va joylashuvi ham turli nashriyotlarda turlicha bo‘lishi mumkin. Ba’zi jurnallarda bo‘limlar raqamlangan, ba’zilarida raqamlanmagan; ba’zilarida “Introduction” o‘rniga “Background” yoki “Motivation” kabi nomlar ishlatilishi mumkin. Bo‘limlarni aniqlash va ularning chegaralarini belgilash metama’lumotlarni ajratishning muhim qismi hisoblanadi, chunki ko‘plab metama’lumot elementlari (masalan, annotatsiya) ma’lum bo‘limlar ichida joylashgan bo‘ladi. Manzoor tomonidan olib borilgan tadqiqotda elektron tezis va dissertatsiyalarni boblarga ajratish uchun pastdan yuqoriga yondashuv taklif qilingan bo‘lib, LaTeX manipulyatsiyasi yordamida mazmunli ma’lumotlar to‘plamlarini yaratish imkoniyati ko‘rsatilgan.[5]

Akademik hujjatlarning formatlari va tuzilishini o‘rganishda ma’lumotlar to‘plamlarining roli alohida ahamiyatga ega. Mashinaviy o‘rganish modellarini

o‘qitish va baholash uchun turli formatlardagi, turli sohalarga oid va turli nashriyotlarning shablonlarida tayyorlangan katta hajmdagi anotatsiyalangan ma’lumotlar to‘plamlari zarur. Li va boshqalar tomonidan yaratilgan DocBank ma’lumotlar to‘plami 500 ming hujjat sahifasini token darajasida anotatsiyalar bilan ta’minlab, hujjat tuzilishini tahlil qilish bo‘yicha tadqiqotlar uchun muhim asos bo‘lib xizmat qiladi.[11] Xuddi shunday, TableBank ma’lumotlar to‘plami Word va LaTeX hujjatlaridan zaif nazorat bilan yaratilgan 417 ming jadvalni o‘z ichiga oladi.[12] Bunday ma’lumotlar to‘plamlari turli formatlardagi akademik hujjatlardan metama’lumotlarni ekstraksiya qilish bo‘yicha modellarni o‘qitish va ularni taqqoslash imkonini beradi.

Akademik hujjatlarning formatlari va tuzilishi metama’lumotlarni avtomatik ekstraksiya qilish tizimlarini ishlab chiqishda hal qiluvchi omil hisoblanadi. PDF formatining keng tarqalganligi va unda strukturaviy ma’lumotlarning yo‘qligi asosiy muammolardan biridir. LaTeX formati mantiqiy strukturani aniq belgilash imkonini bergani bilan, uni qo‘llash sohasi cheklangan. XML va JATS XML esa metama’lumotlarni ifodalashning eng standartlashtirilgan usuli bo‘lsa-da, barcha nashriyotlar tomonidan qo‘llanilmaydi. Turli nashriyotlar va jurnallarning shablonlaridagi xilma-xillilik, skanerlangan hujjatlar muammosi, formatlar o‘rtasidagi konversiya xatolari va hujjatlarning ichki tuzilishidagi farqlar metama’lumotlarni ekstraksiya qilish tizimlariga qo‘yiladigan asosiy talablarni belgilaydi. Keyingi bo‘limlarda ushbu muammolarni hal qilishga qaratilgan mavjud yechimlar va ularning imkoniyatlari batafsil tahlil qilinadi.

Foydalanilgan adabiyotlar

1. Boukhers, Z., & Yang, C. (2025). Comparison of Feature Learning Methods for Metadata Extraction from PDF Scholarly Documents. arXiv preprint arXiv:2501.05082.

2. Z. Boukhers and A. Bouabdallah, "Vision and Natural Language for Metadata Extraction from Scientific PDF Documents: A Multimodal Approach," 2022 ACM/IEEE Joint Conference on Digital Libraries (JCDL), Cologne, Germany, 2022, pp. 1-5.

3. M. Hasan Choudhury, H. R. Jayanetti, J. Wu, W. A. Ingram and E. A. Fox, "Automatic Metadata Extraction Incorporating Visual Features from Scanned Electronic Theses and Dissertations," 2021 ACM/IEEE Joint Conference on Digital Libraries (JCDL), Champaign, IL, USA, 2021, pp. 230-233, doi: 10.1109/JCDL52503.2021.00066.

4. Muntabir Hasan Choudhury, Jian Wu, William A. Ingram, and Edward A. Fox. 2020. A Heuristic Baseline Method for Metadata Extraction from Scanned Electronic Theses and Dissertations. In Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2020 (JCDL '20). Association for Computing Machinery, New York, NY, USA, 515–516. https://doi.org/10.1145/3383583.3398590

5. Manzoor, J.A.: Segmenting electronic theses and dissertations by chapters. MS thesis, Virginia Tech, Computer Science, defended September 23, 2022 (2022).

6. Meng Ling and Jian Chen. 2020. DeepPaperComposer: A Simple Solution for Training Data Preparation for Parsing Research Papers. In Proceedings of the First Workshop on Scholarly Document Processing, pages 91–96. Association for Computational Linguistics.

7. O. Iwashokun and A. Ade-Ibijola, “Parsing of Research Documents into XML Using Formal Grammars,” Applied Computational Intelligence and Soft Computing, vol. 2024, 2024, doi: 10.1155/2024/6671359.

8. Z. Boukhers, N. Beili, T. Hartmann, P. Goswami and M. A. Zafar, "MexPub: Deep Transfer Learning for Metadata Extraction from German

Publications," 2021 ACM/IEEE Joint Conference on Digital Libraries (JCDL), Champaign, IL, USA, 2021, pp. 250-253, doi: 10.1109/JCDL52503.2021.00076.

9. Manzoor, J.A.: Segmenting electronic theses and dissertations by chapters. MS thesis, Virginia Tech, Computer Science, defended September 23, 2022 (2022).

10. Minghao Li, Lei Cui, Shaohan Huang, Furu Wei, Ming Zhou, and Zhoujun Li. 2020. TableBank: Table Benchmark for Image-based Table Detection and Recognition. In Proceedings of the Twelfth Language Resources and Evaluation Conference, pages 1918–1925, Marseille, France. European Language Resources Association.