METAMA’LUMOTLARNI EKSTRAKSIYA QILISH BO‘YICHA MAVJUD YECHIMLAR

Rashid Turg‘unboyev
Qo‘qon davlat universiteti

Annotatsiya: Ushbu maqolada akademik maqolalardan metama’lumotlarni avtomatik ekstraksiya qilish sohasidagi mavjud yechimlar evolyutsion rivojlanish nuqtai nazaridan tahlil qilinadi. Qoidalarga asoslangan tizimlar, an’anaviy mashinaviy o‘rganish modellari (SVM, HMM, Random Forest) va chuqur o‘rganish asosidagi yechimlar (CNN, RNN, LSTM, transformer arxitekturalari) o‘rtasida qiyosiy tahlil o‘tkaziladi. GROBID, CERMINE, Science Parse, Anystyle va ParsCit kabi ochiq kodli tizimlarning imkoniyatlari va cheklovlari batafsil yoritiladi. Tadqiqot natijalariga ko‘ra, qoidalarga asoslangan tizimlar ma’lum formatlarda yuqori aniqlik ko‘rsatsa-da, moslashuvchanligi pastligi bilan ajralib turadi. An’anaviy mashinaviy o‘rganish modellari nisbatan moslashuvchan bo‘lsa-da, sifatli o‘quv ma’lumotlariga bog‘liq. Chuqur o‘rganish asosidagi tizimlar kontekstual va semantik xususiyatlarni o‘rganish qobiliyati tufayli murakkab hollarda barqaror natijalar beradi (masalan, psevdo-kodlarni ajratishda 94.23% aniqlik). Biroq, barcha mavjud tizimlar turli formatlar, skanerlangan hujjatlar va ko‘p tilli maqolalarda universal va barqaror natijalar bera olmaydi.

Kalit so‘zlar: *metama’lumotlarni avtomatik ekstraksiya qilish, akademik maqolalar, qoidalarga asoslangan tizimlar, mashinaviy o‘rganish, SVM, chuqur o‘rganish, CNN, LSTM, GROBID, CERMINE, Science Parse, Anystyle, bibliografik havolalar, skanerlangan hujjatlar, kontekstual tahlil, neyron tarmoqlar, OCR, ko‘p tilli hujjatlar*

Akademik maqolalardan metama’lumotlarni avtomatik ekstraksiya qilish muammosi so‘nggi yillar davomida kompyuter lingvistikasi va axborot tizimlari sohasidagi tadqiqotchilarning diqqat markazida bo‘lib kelmoqda. Ilmiy nashrlar

hajmining keskin o‘sishi va ularning turli formatlarda tarqalishi ushbu muammoni hal qilishga qaratilgan turli yondashuv va tizimlarning paydo bo‘lishiga olib keldi. Ushbu yechimlarni evolyutsion rivojlanish nuqtai nazaridan qaraydigan bo‘lsak, ularni uch asosiy toifaga ajratish mumkin: qoidalarga asoslangan tizimlar, an’anaviy mashinaviy o‘rganish modellari va so‘nggi yillarda rivojlanayotgan chuqur o‘rganish asosidagi yechimlar. Har bir toifa o‘ziga xos afzalliklar va cheklovlarga ega.

Qoidalarga asoslangan tizimlar metama’lumotlarni avtomatik ekstraksiya qilish sohasidagi dastlabki yechimlar hisoblanadi. Ushbu tizimlar inson ekspertlari tomonidan oldindan belgilangan qoidalar va andozalar asosida ishlaydi. Masalan, maqolaning birinchi sahifasidagi katta shriftdagi matn odatda sarlavha deb hisoblanadi, “Abstract” yoki “Annotatsiya” so‘zidan keyin keladigan matn esa annotatsiya sifatida ajratiladi, “References” yoki “Adabiyotlar” bo‘limidagi qatorlar esa bibliografik havolalar deb qabul qilinadi. Qoidalarga asoslangan tizimlarning asosiy afzalligi ularning soddaligi, tushunarli bo‘lishi va kam hisoblash resurslarini talab qilishidir. Ular ma’lum bir jurnal yoki nashriyot formatiga moslashtirilgan bo‘lsa, juda yuqori aniqlik bilan ishlashi mumkin. Hashmi va boshqalar tomonidan taklif qilingan qoidalarga asoslangan yondashuv ESWC 2016 ma’lumotlar to‘plamida CERMINE va GROBID kabi taniqli tizimlardan ustun natija ko‘rsatgan.[1] Xususan, mualliflar tomonidan ishlab chiqilgan usul 22% CERMINE va 9% GROBID dan yuqori ko‘rsatkichga erishgan.

Biroq, qoidalarga asoslangan tizimlarning eng katta kamchiligi ularning moslashuvchanligining pastligidir. Dunyoda minglab turli xil jurnallar mavjud bo‘lib, ularning har biri maqolalarni turlicha formatlaydi. Bir jurnal uchun ishlab chiqilgan qoidalar to‘plami boshqa jurnal maqolalarida umuman ishlamasligi mumkin. Shu sababli, qoidalarga asoslangan tizimlarni yangi formatlar paydo bo‘lganda yoki mavjud formatlar o‘zgarganda doimiy ravishda yangilab borish

zarurati tug‘iladi. Bu esa bunday tizimlarni keng ko‘lamda qo‘llashni qiyinlashtiradi. Qoidalarga asoslangan tizimlarning yana bir cheklovi shundaki, ular asosan hujjatning vizual joylashuviga tayanadi va matnning mazmunini tushunmaydi. Masalan, “Abstract” so‘zi ishlatilmagan, ammo annotatsiya ma’nosidagi matn mavjud bo‘lgan hollarda bunday tizimlar annotatsiyani aniqlay olmasligi mumkin.

An’anaviy mashinaviy o‘rganish modellari metama’lumotlarni ekstraksiya qilish sohasida qoidalarga asoslangan tizimlardan keyingi muhim qadam bo‘ldi. Ushbu modellar matn qatorlarining turli xususiyatlariga (satr uzunligi, shrift o‘lchami, sahifadagi joylashuvi, raqamlar yoki maxsus belgilar mavjudligi, katta-kichik harflar nisbati va boshqalar) asoslanib, ularni turli metama’lumot klasslariga (sarlavha, muallif, annotatsiya, mansublik va boshqalar) ajratishni o‘rganadi. Eng keng tarqalgan modellar qatoriga Support Vector Machine, Hidden Markov Model, Logistic Regression va Random Forest kabi algoritmlar kiradi. Cervantes va boshqalar tomonidan olib borilgan keng qamrovli tadqiqotda SVM algoritmining turli sohalardagi qo‘llanilishi, uning afzalliklari va cheklovlari batafsil tahlil qilingan.[2]

An’anaviy mashinaviy o‘rganish yondashuvining qoidalarga asoslangan tizimlarga nisbatan asosiy afzalligi uning moslashuvchanligidir. Ushbu modellar turli xil formatdagi maqolalar bilan ishlash imkoniyatiga ega, chunki ular o‘quv ma’lumotlari to‘plamida mavjud bo‘lgan turli xil formatlardan umumiy qonuniyatlarni o‘rganadi. Rahnama va boshqalar tomonidan olib borilgan tadqiqotda Eron tezislaridan metama’lumotlarni ajratish uchun SVM asosidagi usul taklif qilingan bo‘lib, mualliflar 91.36% F1-natijaga erishganlar.[3] Ushbu yondashuv ikki bosqichdan iborat – birinchi bosqichda tezisning asosiy qismi chegaralari aniqlanadi, ikkinchi bosqichda qolgan paragraflar SVM yordamida kerakli metama’lumot klasslariga ajratiladi. Beydaghi va boshqalar esa fors tilidagi tezislardan metama’lumotlarni ajratishda ensemble o‘rganish usulini taklif

qilganlar.[4] Ular SVM, KNN va qaror daraxti kabi turli klassifikatorlarni birlashtirib, yagona klassifikator algoritmlariga nisbatan yuqori F1-natijaga erishganlar.

An’anaviy mashinaviy o‘rganish modellarining samarali ishlashi uchun katta hajmdagi va yuqori sifatli, qo‘lda belgilangan o‘quv ma’lumotlari talab qilinadi. O‘quv ma’lumotlarining sifati va hajmi modelning umumlashtirish qobiliyatiga bevosita ta’sir ko‘rsatadi. Ingram va boshqalar tomonidan elektron tezis va dissertatsiyalar uchun mashinaviy o‘rganish modellarini o‘qitish va baholashda foydalanish mumkin bo‘lgan yuqori sifatli ma’lumotlar to‘plamlarini yaratish usullari batafsil bayon qilingan.[5] Mualliflar ma’lumotlarni qo‘lda belgilash va sintetik jarayonlar orqali yaratish usullarini ishlab chiqqanlar. Biroq, bunday yuqori sifatli o‘quv ma’lumotlarini yaratish ko‘p vaqt va mehnat talab qiladigan jarayondir. Bundan tashqari, an’anaviy mashinaviy o‘rganish modellari ko‘pincha maqolaning vizual joylashuviga tayanadi va matnning mazmunini to‘liq tushunmaydi. Bu esa murakkab yoki nostandart hollarda xatolikka olib kelishi mumkin. Xususan, turli tillardagi maqolalar, murakkab sarlavhalar yoki bir nechta mualliflar va ularning mansubligi haqidagi ma’lumotlarni aniqlashda an’anaviy mashinaviy o‘rganish modellari ba’zan yetarli darajada samarali bo‘la olmaydi.

Chuqur o‘rganish asosidagi yechimlar so‘nggi yillarda metama’lumotlarni ekstraksiya qilish sohasida muhim yutuqlarga erishdi. Konvolyutsion neyron tarmoqlar, takrorlanuvchi neyron tarmoqlar, uzun qisqa muddatli xotira va transformer arxitekturalari asosidagi modellar matnning kontekstual xususiyatlarini chuqur o‘rganish imkonini beradi. Safder va boshqalar tomonidan olib borilgan tadqiqotda algoritmik metama’lumotlarni ajratish uchun chuqur neyron tarmoqlarga asoslangan usul taklif qilingan.[6] Mualliflar 93 mingdan ortiq matn qatorlari ustida o‘tkazilgan tajribada mazmun, shrift uslubi va strukturaga asoslangan 60 ta yangi xususiyatni kiritgan holda algoritmik psevdo-kodlarni ajratish usulini ishlab

chiqqanlar. Taklif qilingan usul 93.32% F1-natijaga erishib, mavjud eng yaxshi texnikalarni 28% ga ortda qoldirgan. Algoritm bilan bog‘liq jumlalarni ajratishda esa chuqur neyron tarmoqlar 78.5% aniqlikka erishgan bo‘lib, bu qoidalarga asoslangan modeldan 28% va SVM modelidan 16% yuqori ko‘rsatkichdir.

Chuqur o‘rganish modellarining muhim afzalligi shundaki, ular matnning kontekstual va semantik xususiyatlarini o‘z-o‘zidan o‘rgana oladi. Bu ularga an’anaviy mashinaviy o‘rganish modellariga qaraganda murakkabroq va nostandart hollarda ham nisbatan barqaror natijalar berish imkonini beradi. Raghavendra Nayaka va Ranjan tomonidan olib borilgan tadqiqotda ensemble CNN va BiLSTM texnikasidan foydalangan holda ilmiy hujjatlardan metama’lumotlarni ajratish samaradorligi oshirilgan.[7] Ushbu yondashuv psevdo-kodlarni ajratishda 94.23% klassifikatsiya aniqligiga, algoritm bilan bog‘liq jumlalarni ajratishda esa 82% aniqlikka erishgan. Bu natijalar qoidalarga asoslangan (23.5%) va SVM (21.5%) usullaridan sezilarli darajada yuqoridir.

Amaliyotda keng qo‘llaniladigan ochiq kodli tizimlar orasida GROBID, CERMINE, Science Parse, Anystyle va ParsCit kabi vositalarni alohida ajratish mumkin. GROBID nemis axborot texnologiyalari tadqiqot markazi tomonidan ishlab chiqilgan bo‘lib, u mashinaviy o‘rganish asosida ishlaydi va PDF formatidagi ilmiy maqolalardan metama’lumotlarni (sarlavha, mualliflar, annotatsiya, adabiyotlar) ajratib olishga ixtisoslashgan. CERMINE esa Polsha fanlar akademiyasi tomonidan ishlab chiqilgan bo‘lib, u modulli arxitekturaga ega va turli mashinaviy o‘rganish usullaridan foydalanadi. Tkaczyk va boshqalar tomonidan taqdim etilgan CERMINE tizimi 18-jild, 4-son, 317-335-betlarda batafsil tavsiflangan.[8] Ushbu tizimning o‘rtacha F1-natijasi 77.5% ni tashkil etadi va u ochiq kodli litsenziya ostida taqdim etilgan.

Ushbu tizimlarning qiyosiy tahlili ularning turli formatdagi va turli sohalarga oid maqolalardagi samaradorligini baholashga qaratilgan bir qator tadqiqotlarda

o‘tkazilgan. Meuschke va boshqalar tomonidan olib borilgan keng qamrovli tadqiqotda o‘nta bepul vositaning (GROBID, CERMINE, Science Parse, Adobe Extract va boshqalar) akademik PDF hujjatlardan metama’lumotlar, bibliografik havolalar, jadvallar va boshqa tarkibiy elementlarni ekstraksiya qilish samaradorligi qiyosiy tahlil qilingan.[9] Tadqiqot natijalari shuni ko‘rsatadiki, GROBID metama’lumotlar va adabiyotlarni ajratish bo‘yicha eng yaxshi natijalarni ko‘rsatgan, undan keyin CERMINE va Science Parse joylashgan. Jadvallarni ajratish bo‘yicha esa Adobe Extract boshqa vositalardan ustun kelgan, ammo uning samaradorligi boshqa tarkibiy elementlarga nisbatan ancha past. Barcha vositalar ro‘yxatlar, kolontitullar va tenglamalarni ajratishda qiyinchiliklarga duch kelgan. Cioffi va Peroni tomonidan olib borilgan tadqiqotda esa bibliografik havolalarni ajratish va tahlil qilish bo‘yicha yettita vosita (Anystyle, Cermine, ExCite, Grobid, Pdfssa4met, Scholarcy va Science Parse) 56 ta PDF maqolalar korpusida qiyosiy baholangan.[10] Anystyle umumiy natijalar bo‘yicha eng yaxshi ko‘rsatkichni qayd etgan, undan keyin Cermine joylashgan. Biroq, ayrim fan sohalarida va ayrim vazifalar bo‘yicha boshqa vositalar yaxshiroq natijalar ko‘rsatgan.

Mavjud yechimlarning asosiy muammolaridan biri ularning turli xil formatdagi va tuzilishdagi maqolalarda barqaror va yuqori sifatli natija bera olmasligidir. Bir vosita ma’lum bir nashriyot yoki jurnal formatida yuqori aniqlik ko‘rsatsa, boshqa formatda keskin pasayib ketishi mumkin. Bu muammo, ayniqsa, nemis ijtimoiy fanlari kabi turli xil shablonlar keng qo‘llaniladigan sohalarda keskin namoyon bo‘ladi. Boukhers va boshqalar tomonidan ta’kidlanganidek, nemis nashrlarida metama’lumotlarning tartibi, mazmuni, joylashuvi va shrift o‘lchami nashrlar orasida juda katta farq qiladi, bu esa an’anaviy NLP usullarining bunday nashrlardan metama’lumotlarni aniq ekstraksiya qilishda yetarli samaradorlikka erisha olmasligiga olib keladi.[11] Yana bir muhim muammo - skanerlangan hujjatlar bilan ishlashdagi qiyinchilikdir. GROBID va CERMINE kabi tug‘ma

hujjatlar uchun ishlab chiqilgan tizimlar skanerlangan elektron tezis va dissertatsiyalarda ko‘pincha ishlamaydi yoki past sifatli natijalar beradi. Choudhury va boshqalar tomonidan olib borilgan tadqiqotda CRF modeliga vizual xususiyatlarni qo‘shish orqali skanerlangan hujjatlardan metama’lumotlarni ekstraksiya qilish samaradorligini oshirish mumkinligi ko‘rsatilgan.[12]

Metama’lumotlarni ekstraksiya qilish bo‘yicha mavjud yechimlar qoidalarga asoslangan tizimlardan tortib, an’anaviy mashinaviy o‘rganish modellari va zamonaviy chuqur o‘rganish asosidagi tizimlargacha bo‘lgan keng spektrni qamrab oladi. Har bir toifa o‘ziga xos afzalliklar va cheklovlarga ega. Qoidalarga asoslangan tizimlar ma’lum bir formatga moslashtirilganda yuqori aniqlik ko‘rsatsa-da, ularning moslashuvchanligi past va doimiy yangilab borishni talab qiladi. An’anaviy mashinaviy o‘rganish modellari nisbatan moslashuvchan bo‘lsa-da, ularning samaradorligi o‘quv ma’lumotlarining sifati va hajmiga bog‘liq. Chuqur o‘rganish asosidagi tizimlar kontekstual va semantik xususiyatlarni o‘rganish qobiliyati tufayli murakkab va nostandart hollarda ham nisbatan barqaror natijalar beradi. Biroq, barcha mavjud tizimlar turli xil formatdagi va tuzilishdagi maqolalarda, ayniqsa skanerlangan hujjatlar va ko‘p tilli maqolalarda, yetarli darajada universal va barqaror natija bera olmaydi. Ushbu muammolarni hal qilishda katta til modellari asosidagi yondashuvlar yangi imkoniyatlar eshigini ochmoqda, ular keyingi bo‘limda batafsil tahlil qilinadi.

Foydalanilgan adabiyotlar ro‘yxati

1. Hashmi A.M., Afzal M.T., Rehman S.u. Rule Based Approach to Extract Metadata from Scientific PDF Documents // Proceedings of the 2020 5th International Conference on Innovative Technologies in Intelligent Systems and Industrial Applications (CITISIA). – Sydney, Australia, 2020. – P. 1-4. – DOI: 10.1109/CITISIA50690.2020.9371784.

2. Cervantes J., Garcia-Lamont F., Rodríguez-Mazahua L., Lopez A. A Comprehensive Survey on Support Vector Machine Classification: Applications, Challenges and Trends // Neurocomputing. – 2020. – Vol. 408. – P. 189–215.

3. Rahnama M., Hasheminejad S.M.H., Nasiri J.A. Automatic Metadata Extraction From Iranian Theses and Dissertations // Proceedings of the 2020 6th Iranian Conference on Signal Processing and Intelligent Systems (ICSPIS). – Mashhad, Iran, 2020. – P. 1-5. – DOI: 10.1109/ICSPIS51611.2020.9349570.

4. Beydaghi E., Rahnama M., Nasiri J.A. Ensemble Approach for Metadata Extraction in Persian Theses // Proceedings of the 2020 6th Iranian Conference on Signal Processing and Intelligent Systems (ICSPIS). – Mashhad, Iran, 2020. – P. 1-5. – DOI: 10.1109/ICSPIS51611.2020.9349571.

5. Ingram W.A., Wu J., Kahu S.Y. et al. Building Datasets to Support Information Extraction and Structure Parsing from Electronic Theses and Dissertations // International Journal on Digital Libraries. – 2024. – Vol. 25. – P. 175-196. – DOI: 10.1007/s00799-024-00395-4.

6. Safder I., Hassan S.-U., Visvizi A., Noraset T., Nawaz R., Tuarob S. Deep Learning-Based Extraction of Algorithmic Metadata in Full-Text Scholarly Documents // Information Processing and Management. – 2020. – Vol. 57, No. 6. – Art. 102269.

7. Raghavendra Nayaka P., Ranjan R. An Efficient Framework for Metadata Extraction over Scholarly Documents Using Ensemble CNN and BiLSTM Technique // Proceedings of the 2023 2nd International Conference for Innovation in Technology (INOCON). – Bangalore, India, 2023. – P. 1-9. – DOI: 10.1109/INOCON57975.2023.10101029.

8. Tkaczyk D., Szostek P., Fedoryszak M., Dendek P.J., Bolikowski L. CERMINE: Automatic Extraction of Structured Metadata from Scientific Literature

// International Journal on Document Analysis and Recognition. – 2015. – Vol. 18, No. 4. – P. 317-335.

9. Meuschke N., Jagdale A., Spinde T., Mitrović J., Gipp B. A Benchmark of PDF Information Extraction Tools Using a Multi-Task and Multi-Domain Evaluation Framework for Academic Documents // Information for a Better World: Normality, Virtuality, Physicality, Inclusivity. Proceedings of iConference 2023 / ed. by I. Sserwanga et al. – Cham: Springer, 2023. – Vol. 13972 (Lecture Notes in Computer Science). – P. 383-405.

10. Cioffi A., Peroni S. Structured References from PDF Articles: Assessing the Tools for Bibliographic Reference Extraction and Parsing // International Conference on Theory and Practice of Digital Libraries. – Cham: Springer International Publishing, 2022.

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

12. Hasan Choudhury M., Jayanetti H.R., Wu J., Ingram W.A., Fox E.A. Automatic Metadata Extraction Incorporating Visual Features from Scanned Electronic Theses and Dissertations // arXiv e-prints. – 2021. – arXiv:2107.