ISNN 3030-3907

2026/4
VOLUME 7

DEVELOPMENT OF SCIENCE

ILMIY JURNAL



Development of science

Ilmiy jurnal

2026/4 VOLUME 7

ISSN 3030 -3907

Ilmiy jurnal OAK rayosatining 2024 yil 27 sentabrdagi 361-son qaroriga asosan 02.00.00 kimyo fanlari hamda 2025 yil 12 fevral 367- son qaroriga asosan 05.00.00 texnika fanlari bo‘yicha falsafa doktori (PhD) va fan doktori (DSc) ilmiy darajasiga talabgorlarning dissertatsiya ishlari yuzasidan dissertatsiyalari asosiy ilmiy natijalarini chop etish tavsiya etilgan milliy ilmiy nashrlar ro‘yxatiga kiritilgan.

BOSH MUHARRIR:
Temirov Alisher Hoshim o‘g‘li – texnika fanlari falsafa doktori (PhD),Dotsent.
TAHRIRIYAT KENGASH RAISI:
Fozilov Sadriddin Fayzullayevich – texnika fanlari doktori, professor
TAHRIRIYAT KENGASH A’ZOLARI

TEXNIKA FANLARI/ TECHNICAL SCIENCES/ ТЕХНИЧЕСКИЕ НАУКИ

Bozorov G‘ayrat Rashidovich
Buxoro Davlat Texnika Universiteti, professor
Abdurahmonov Olim Rustamovich
Buxoro Davlat Texnika Universiteti, professor
Adizov Bobur Zamirovich
O‘zbekiston Respublikasi Fanlar akademiyasi Umumiy va noorganik kimyo instituti, professor
Hayitov Ruslan Rustamjonovich
Buxoro Davlat Texnika Universiteti, professor
Jumayev Qayum Karimovich
Buxoro Davlat Texnika Universiteti, professor.
Berdiyev Dorob Murodovich
Islom Karimov nomidagi Toshkent Davlat Texnika Universiteti, professor
Maxmudov Rafiq Amonovich
Buxoro Davlat Texnika Universiteti, professor.
Tilloyev Lochin Ismatilloyevich
Buxoro Davlat Texnika Universiteti, dotsent.
Raxmonov Qahramon Sanakulovich
Buxoro Davlat Texnika Universiteti, dotsent.
Barakayev Nusratullo Rajabovich
Islom Karimov nomidagi Toshkent Davlat Texnika Universiteti, professor.
Jo‘rayev Xayrullo Fayziyevich
Buxoro Davlat Texnika Universiteti, professor.
Panoyev Erali Rajabboyevich
Buxoro Davlat Texnika Universiteti, dotsent.
Ataullayev Sherzod Nabiullyevich
Buxoro Davlat Texnika Universiteti, dotsent.
Murodov Malik Negmurodovich
Buxoro Davlat Texnika Universiteti, dotsent.
Axmedova Ozoda Bahronovna
Islom Karimov nomidagi Toshkent Davlat Texnika Universiteti.
Safarov Bahri Jumayevich
Buxoro Davlat Texnika Universiteti, dotsent.
Xaydarov Axtam Amonovich
Buxoro Davlat Texnika Universiteti, dotsent.
Safarov Jasur Alijon o‘g‘li
Buxoro Davlat Texnika Universiteti, dotsent.
Асадова Раъно Дилмуратовна
Тошкент Давлат техника университети доценти

KIMYO FANLARI/ CHEMICAL SCIENCES/ ХИМИЧЕСКИЕ НАУКИ

Axmedov Voxid Nizomovich
Buxoro Davlat Texnika Universiteti, professor
Umarov Baqo Bafoyevich
Buxoro Davlat Universiteti, professor
Kodirov Abduaxad Abduraximovich
Qarshi davlat Universiteti, professor
Qurbonov Mingniqul Jumagulovich
Qarshi davlat Universiteti, dotsent
Turemuratov Sharibay Naurizbaevich
O‘zbekiston Respublikasi Fanlar akademiyasi Qoraqalpog‘iston bo‘limi Bosh ilmiy kotibi, professor
Tursunov Kaxor Shonazarovich
Qarshi davlat Universiteti, professor
Do‘stov Hamro
Buxoro Davlat Texnika Universiteti, professor
Panoyev Nodir Shavkatovich
Buxoro Davlat Texnika Universiteti, dotsent.
Olimov Bobur Bahodir o‘g‘li
Buxoro Davlat Texnika Universiteti, dotsent.
Kasimov Sherzod Abduazimovich
Termiz Davlat Universiteti, professor.
Aliqulov Rustam Valiyevich
Termiz Davlat Universiteti, professor
Tropov Xamza Tursunovich
Samarqand Davlat Universiteti, professor

IQTISODIYOT FANLARI/ ECONOMICS/ ЭКОНОМИЧЕСКИЕ НАУКИ

Nizamov Aslitdin Badreddinovich
Buxoro Davlat Texnika Universiteti, professor.
Boboyev Akmal Choriyevich
Buxoro Davlat Texnika Universiteti, professor.
Ergashxojayeva Shaxnoza Djasurovna
Toshkent davlat iqtisodiyot universiteti, professor
Saidaxmedova Nodira Ilxomovna
Toshkent davlat iqtisodiyot universiteti, professor
Muxsinov Bekzod Toxirovich
Buxoro Davlat Texnika Universiteti, dotsent.
Shakarov Qulmat Ashirovich
Renessans ta'lim universiteti,dotsent.
Jumaeva Zulfiya Qayumovna
Buxoro Davlat Texnika Universiteti, dotsent.
Norova Salomat Yusupovna
Buxoro Davlat Texnika Universiteti, dotsent.
Avezova Shaxnoza Maxmudjonovna
Buxoro Davlat Texnika Universiteti, dotsent.
Jalilov Jamshid G‘anijonovich
Toshkent davlat iqtisodiyot universiteti, dotsent
Xasanova Gulruh Djumanazarovna
Buxoro Davlat Texnika Universiteti, dotsent.

PEDAGOGIKA FANLARI/ PEDAGOGICAL SCIENCES / ПЕДАГОГИЧЕСКИЕ НАУКИ

Siddiqova Sadoqat G‘afforovna
Buxoro Davlat Texnika Universiteti, dotsent.
Zoyirov Erkin Xalilovich
Buxoro Davlat Texnika Universiteti, DSc, professor
Gaffarov Feruz Hasanovich
Puchon Universiteti, *professor*
Ruziyev Davron Yuldashevich
Buxoro Davlat Universiteti, professor
Muradova Firuza Rashidovna
Buxoro Davlat Texnika Universiteti, professor.
Yunusova Gulandon Samiyevna
Buxoro Davlat Texnika Universiteti, professor.
To‘xtayeva Zebo Sharifovna
Buxoro Davlat Texnika Universiteti, professor.
Navro‘zova Gulchixra Nigmatovna
Buxoro Davlat Texnika Universiteti, dotsent.
Uzoqova Laylo Polvonovna
Buxoro muhandislik – texnologiya instituti, dotsent.
To‘raqulova Marjona Qiyom qizi
Buxoro Davlat Texnika Universiteti, dotsent.
Muradova Zarina Rashidovna
Buxoro Davlat Texnika Universiteti, dotsent.
Murodov Sanjar Aslonovich
Buxoro Davlat Texnika Universiteti, dotsent.
Shodiyev Jahongir Jo‘raqulovich
Buxoro Davlat Texnika Universiteti, dotsent
Gaffarov Laziz Hasanovich
Buxoro muhandislik – texnologiya instituti, dotsent.

Tahririyat a’zolarning to‘liq ro‘yxati jurnalning www.devos.uz saytida keltirilgan.

Tahririyat hay’ati a’zolarining tarkibini o‘zgartirish hamda qo‘shimcha kiritish yoki tahririyat a’zoligidan chiqarish bosh muharrir tomonidan amalga oshiriladi.

ISSN 3030-3907

Development of science ilmiy jurnali 2024 yil 15 aprel kuni №256445 sonli guvohnoma bilan Davlat ro‘yxatidan o‘tgan.

To‘plam elektron ko‘rinishda (pdf formatida) mualliflarga taqdim etiladi. To‘plamga kiritiladigan maqolalarning mazmuni, ulardagi statistik ma’lumotlar va me’yoriy hujjatlarning to‘g‘riligiga hamda har qanday fikr - mulohazalarga mualliflarning o‘zlari mas’ul hisoblanadilar. Talablarga javob bermaydigan maqolalar to‘plamga kiritilmaydi. Tashkiliy qo‘mita maqola matnini qisqartirish, qisman tuzatish kiritish, seksiyalarga joylashtirish huquqiga ega.

ISSN 3030-3907

AKADEMIK MAQOLALARDAN METAMA'LUMOTLARNI EKSTRAKSIYA QILISH: MUAMMO TAHLILI VA MAVJUD YECHIMLAR

Rashid Muxtorovich Turgunbayev
Qo'qon davlat universiteti dotsenti
Qo'qon shahar, O'zbekiston
E-mail: atmdsmi@gmail.com

Annotatsiya. Akademik maqolalardan metama'lumotlarni avtomatik ekstraksiya qilish ilmiy nashrlarni indeksatsiyalash va qidirish uchun muhim ahamiyatga ega. Ushbu maqolada metama'lumot turlari, akademik hujjat formatlarining (PDF, LaTeX, XML) ta'siri hamda mavjud yechimlar - qoidalarga asoslangan tizimlar, an'anaviy mashinaviy o'rganish va chuqur o'rganing modellari tahlil qilinadi. GROBID, CERMINE va Anystyle kabi tizimlarning cheklovlari ko'rsatilgan. So'nggi muammo sifatida ko'p tillilik va o'zbek tili uchun tizimlarning yetishmasligi aniqlangan.

Kalit so'zlar: metama'lumot, akademik maqolalar, metama'lumotlarni ekstraksiya qilish, mashinaviy o'rganish, chuqur o'rganish, PDF format

Abstract. Automatic extraction of metadata from academic articles is of great importance for indexing and searching scientific publications. This article analyzes the types of metadata, the impact of academic document formats (PDF, LaTeX, XML), and existing solutions - rule-based systems, traditional machine learning and deep learning models. The limitations of systems such as GROBID, CERMINE and Anystyle are shown. The last problem is the lack of systems for multilingualism and the Uzbek language.

Keywords: metadata, academic articles, metadata extraction, machine learning, deep learning, PDF format

Аннотация. Автоматическое извлечение метаданных из научных статей имеет большое значение для индексирования и поиска научных публикаций. В данной статье анализируются типы метаданных, влияние форматов научных документов (PDF, LaTeX, XML) и существующие решения - системы, основанные на правилах, традиционные модели машинного обучения и глубокого обучения. Показаны ограничения таких систем, как GROBID, CERMINE и Anystyle. Последняя проблема - отсутствие систем для многоязычия и узбекского языка.

Ключевые слова: метаданные, научные статьи, извлечение метаданных, машинное обучение, глубокое обучение, формат PDF

KIRISH.

ISSN 3030-3907

Zamonaviy ilm-fan taraqqiyotining asosiy poydevori bo‘lgan akademik maqolalar tadqiqotchilar o‘rtasida bilim almashishning eng muhim vositasidir. Har yili dunyo miqyosida millionlab ilmiy maqolalar nashr etiladi va ularning soni muttasil o‘sib bormoqda. Biroq bunday katta hajmdagi ma’lumotlar to‘plamida kerakli bilimlarni topish, ularni tizimlashtirish va qayta ishlash tobora murakkab masalaga aylanib bormoqda. Aynan shu nuqtada metama’lumotlar tushunchasi va ularni avtomatik ravishda ekstraksiya qilish muammosi dolzarb ahamiyat kasb etadi. Metama’lumotlar - bu “ma’lumotlar haqidagi ma’lumotlar” bo‘lib, ular akademik maqolalarning mazmunini tushunish, qidirish va boshqarish uchun kalit vazifasini o‘taydi.[1] Sarlavha, mualliflar ro‘yxati, annotatsiya, kalit so‘zlar, nashr yili, jurnal nomi, iqtiboslar va foydalanilgan adabiyotlar - bularning barchasi metama’lumotlarning muhim tarkibiy qismlaridir.

Metama’lumotlarning to‘g‘ri va to‘liq aniqlanishi ilmiy maqolalarni samarali indeksatsiyalash, ularni ilmiy ma’lumotlar bazalarida qidirish va tadqiqotchilar o‘rtasida ularning ko‘rinishini ta’minlash uchun hal qiluvchi ahamiyatga ega. Shuningdek, metama’lumotlar ilmiy aloqalar tahlili, bibliometrik tadqiqotlar va fanning rivojlanish tendensiyalarini o‘rganishda ham asosiy manba hisoblanadi. Metama’lumotlarni qo‘lda kiritish va tekshirish nafaqat ko‘p vaqt talab qiladigan, balki inson omili bilan bog‘liq xatoliklarga ham yo‘l qo‘yadigan jarayondir.[2] Shu sababli, akademik maqolalardan metama’lumotlarni avtomatlashtirilgan tizimlar yordamida ekstraksiya qilish masalasi bugungi kunda axborot texnologiyalari va kompyuter lingvistikasi sohasidagi eng muhim tadqiqot yo‘nalishlaridan biriga aylangan.

Ushbu maqolada akademik maqolalardan metama’lumotlarni ekstraksiya qilish muammosining nazariy asoslari, mavjud yechimlarning evolyutsiyasi hamda ularning afzalliklari va cheklovlari tahlil qilinadi. Asosiy e’tibor metama’lumot turlari, akademik hujjatlarning formatlari hamda qoidalarga asoslangan, an’anaviy mashinaviy o‘rganish va chuqur o‘rganish yondashuvlariga qaratilgan. Maqola ushbu sohadagi bo‘shliqlarni aniqlash va kelgusidagi tadqiqotlar uchun yo‘nalishlarni belgilashga xizmat qiladi.

ADABIYOTLAR SHARHI VA METODLAR

Metama’lumot tushunchasi va uning akademik nashrlardagi roli sohasida keng qamrovli tadqiqotlar olib borilgan. Metama’lumot, eng sodda ta’rifda, “ma’lumotlar haqidagi ma’lumot” deb izohlanadi.[3] Ilmiy maqolalar kontekstida metama’lumotlar odatda uchta asosiy guruhga bo‘linadi. Birinchi guruh identifikatsiya bilan bog‘liq bo‘lib, maqolaning sarlavhasi, mualliflar ismi, ularning ilmiy darajalari va mansubligi, nashr etilgan jurnal nomi, nashr yili, jild va sahifa raqamlari hamda DOI kabi unikal identifikatorlarni o‘z ichiga oladi.

ISSN 3030-3907

Sarlavha maqola mazmunining qisqacha ifodasi bo‘lib, uni qidirish tizimlarida topishning asosiy kalitlaridan biri hisoblanadi. Mualliflar haqidagi ma’lumotlar ilmiy ishlarning mualliflik huquqi, iqtiboslar tahlili va ilmiy hamkorlikni o‘rganishda muhim rol o‘ynaydi. Ikkinchi guruh maqolaning mazmunini ifodalovchi metama’lumotlardan iborat bo‘lib, bularga annotatsiya, kalit so‘zlar, maqolaning asosiy qismlari haqidagi ma’lumotlar kiradi. Annotatsiya maqolaning qisqacha mazmunini ochib beradi va ko‘plab ilmiy ma’lumotlar bazalarida maqolani indeksatsiyalashda asosiy manba hisoblanadi. Uchinchi guruh metama’lumotlar maqolaning boshqa ilmiy ishlar bilan bog‘liqligini aks ettiradi, jumladan foydalanilgan adabiyotlar, iqtiboslar va ularning konteksti.

Rasuli va boshqalar tomonidan olib borilgan tadqiqotda dissertatsiyalarning metama’lumotlar to‘liqligi va ularning ko‘rishlar va yuklab olishlar soni o‘rtasida ijobiy korrelyatsiya mavjudligi empirik tarzda isbotlangan. Xususan, annotatsiyaning uzunligi va metama’lumot maydonlarining to‘liqligi dissertatsiyalarning ko‘rinishiga sezilarli darajada ta’sir qilishi aniqlangan. Bu esa metama’lumotlarning nafaqat texnik jihatdan to‘g‘ri ajratib olinishi, balki ularning mazmun jihatdan ham to‘liq va ifodali bo‘lishi muhimligini ko‘rsatadi. Metama’lumotlarni tashkil etish va ularni standartlashtirish masalasi xalqaro miqyosda bir qator tashkilotlar tomonidan amalga oshirilmoqda. Dublin Core Metadata Initiative eng keng tarqalgan metama’lumot standartlaridan biri bo‘lib, u o‘n beshta asosiy elementni o‘z ichiga oladi. Ilmiy nashrlar uchun esa JATS XML standarti keng qo‘llaniladi, u maqolalarning barcha strukturaviy elementlarini maxsus teglar bilan belgilash imkonini beradi.

Akademik maqolalardan metama’lumotlarni avtomatik ekstraksiya qilish tizimlarini ishlab chiqishda eng muhim omillardan biri hujjatlarning formatlari va ularning ichki tuzilish xususiyatlarini chuqur tushunishdir. PDF formati bugungi kunda akademik nashrlar sohasida eng keng tarqalgan va eng ko‘p qo‘llaniladigan format hisoblanadi. Uning asosiy afzalligi shundaki, u hujjatning vizual ko‘rinishini shriftlar, rasmlar, jadvallar va sahifalarning joylashuvini 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. PDF formatidagi hujjatlarda matn va boshqa elementlar asosan vizual joylashuv asosida ifodalanadi, ya’ni ularning mantiqiy strukturasi - qaysi qism sarlavha, qaysi qism muallif ismi ekanligi - haqida hech qanday ma’lumot mavjud emas.[4] PDF formatining yana bir muhim xususiyati uning ikki xil turda mavjudligidir: tug‘ma PDF va skanerlangan PDF. Tug‘ma PDF

ISSN 3030-3907

hujjatlar dastlab elektron shaklda yaratilgan bo‘lib, ulardagi matnlar odatda to‘g‘ridan-to‘g‘ri ajratib olish 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.[5]

LaTeX formati ilmiy-texnik hujjatlarni tayyorlash uchun maxsus ishlab chiqilgan matnli formatdir. Uning asosiy afzalligi shundaki, unda hujjatning mantiqiy strukturasi maxsus buyruqlar yordamida aniq belgilanadi. 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. Biroq LaTeX hujjatlari asosan aniq fanlar sohasidagi maqolalar uchun qo‘llaniladi va barcha nashriyotlar tomonidan qabul qilinmaydi.[6] XML formati metama’lumotlarni ifodalashning eng tuzilgan va standartlashtirilgan usuli hisoblanadi. Ilmiy nashrlar sohasida eng keng tarqalgan XML standarti JATS hisoblanadi. Ushbu format ma’lumotlarni mashinalar tomonidan o‘qilishi va qayta ishlanishini maksimal darajada osonlashtiradi. Ko‘plab yirik nashriyotlar va ilmiy ma’lumotlar bazalari o‘z maqolalarini JATS XML formatida saqlaydi va tarqatadi. Biroq barcha nashriyotlar va jurnallar o‘z maqolalarini XML formatida taqdim etavermaydi, shuning uchun metama’lumotlarni ajratish tizimlari asosan PDF formatidagi maqolalar bilan ishlashga majbur.

Akademik hujjatlarning formatlari bilan bog‘liq yana bir muhim jihat - ularning strukturaviy murakkabligi va xilma-xilligidir. Har bir nashriyot va 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. 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.

Metama’lumotlarni avtomatik ekstraksiya qilish sohasidagi yechimlarni uch asosiy bosqichda rivojlangan deb qarash mumkin. Dastlabki yechimlar qoidalarga asoslangan tizimlar edi.[7] Ushbu tizimlar maqolalarning strukturasi haqidagi

ISSN 3030-3907

oldindan belgilangan bilimlarga tayanib ishlaydi. Masalan, maqolaning birinchi sahifasidagi katta shriftdagi matn odatda sarlavha, “Abstract” so‘zidan keyin keladigan parcha annotatsiya, “References” bo‘limidagi qatorlar esa iqtiboslar deb hisoblanadi. Qoidalarga asoslangan tizimlarning asosiy afzalligi ularning tez ishlashi va tushunarli bo‘lishidir. 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. 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.

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, jumladan satr uzunligi, shrift o‘lchami, sahifadagi joylashuvi, raqamlar yoki maxsus belgilar mavjudligi, katta-kichik harflar nisbati kabi ko‘rsatkichlarga asoslanib, ularni turli metama’lumot klasslariga 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. 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 foizlik F1-natijaga erishganlar. Beydaghi va boshqalar esa fors tilidagi tezislardan metama’lumotlarni ajratishda ensemble o‘rganish usulini taklif qilganlar.

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

ISSN 3030-3907

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. 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.

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. Mualliflar 93 mingdan ortiq matn qatorlari ustida o‘tkazilgan tajribada mazmun, shrift uslubi va strukturaga asoslangan oltmishta yangi xususiyatni kiritgan holda algoritmik psevdo-kodlarni ajratish usulini ishlab chiqqanlar. Taklif qilingan usul 93.32 foizlik F1-natijaga erishib, mavjud eng yaxshi texnikalarni 28 foizga ortda qoldirgan. 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.

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 ajratib olishga ixtisoslashgan. CERMINE esa Polsha fanlar akademiyasi tomonidan ishlab chiqilgan bo‘lib, u modulli arxitekturaga ega va turli mashinaviy o‘rganish usullaridan foydalanadi. Meuschke va boshqalar tomonidan olib borilgan keng qamrovli tadqiqotda o‘nta bepul vositaning akademik PDF hujjatlardan metama’lumotlar, bibliografik havolalar, jadvallar va boshqa tarkibiy elementlarni ekstraksiya qilish samaradorligi qiyosiy tahlil qilingan. 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. Cioffi va Peroni tomonidan olib borilgan tadqiqotda esa bibliografik havolalarni ajratish va tahlil qilish bo‘yicha yettita vosita ellik oltita PDF maqolalar korpusida qiyosiy baholangan. Anystyle umumiy natijalar bo‘yicha eng yaxshi ko‘rsatkichni qayd

ISSN 3030-3907

etgan, undan keyin Cermine joylashgan. Biroq ayrim fan sohalarida va ayrim vazifalar bo‘yicha boshqa vositalar yaxshiroq natijalar ko‘rsatgan.

NATIJALAR.

Adabiyotlar va mavjud tizimlar tahlili natijasida bir qator muhim empirik topilmalarga erishildi. Metama’lumotlarning to‘liqligi va sifati ilmiy maqolalarning topiluvchanligi va ta’siriga bevosita ta’sir ko‘rsatadi. Rasuli va boshqalar dissertatsiyalarda metama’lumot maydonlari to‘liqligi bilan ko‘rish va yuklab olishlar soni o‘rtasida ijobiy korrelyatsiyani empirik tarzda isbotlagan. Bu esa metama’lumotlarni ajratish tizimlarining yuqori aniqlikda ishlashi zarurligini ko‘rsatadi.

PDF formatining keng tarqalganligi va unda strukturaviy ma’lumotning yo‘qligi asosiy texnik to‘siq hisoblanadi. Tug‘ma PDF uchun GROBID va CERMINE kabi vositalar yetmishdan sakson besh foizgacha bo‘lgan F1 oralig‘ida ishlaydi. Biroq skanerlangan PDF hujjatlar uchun bu ko‘rsatkich qirqdan oltmish foizgacha tushib ketadi. Boukhers va boshqalar nemis ijtimoiy fanlari nashrlarida an’anaviy NLP usullarining samaradorligi keskin pasayishini qayd etgan va kompyuter ko‘rish yondashuvlarini taklif qilgan.

Qoidalarga asoslangan tizimlar ma’lum jurnallar uchun yuqori aniqlik ko‘rsatsa-da, ularning moslashuvchanligi past. Hashmi va boshqalar qoidalarga asoslangan usul ESWC 2016 korpusida CERMINE dan yigirma ikki foiz va GROBID dan to‘qqiz foiz yuqori natijaga erishgan. Biroq shu tizim boshqa jurnal korpuslarida yigirma-o‘ttiz foizgacha pasaygan. Bu holat qoidalarga asoslangan tizimlarning universal emasligini ko‘rsatadi.

An’anaviy mashinaviy o‘rganish modellari, xususan SVM va ensemble usullar, qoidalarga asoslangan tizimlarga nisbatan moslashuvchanroqdir.[8] Rahnama va boshqalar SVM asosida 91.36 foizlik F1 natijasiga erishgan. Beydaghi va boshqalar ensemble yondashuv bilan fors tilidagi tezislarda yagona klassifikatorlardan uch-besh foiz yuqori natija olgan. Biroq ushbu modellarning samaradorligi anotatsiyalangan o‘quv ma’lumotlari hajmi va sifatiga kuchli bog‘liqdir.

Chuqur o‘rganish modellari kontekstual va semantik tushunish qobiliyati tufayli murakkab va nostandart hollarda eng yaxshi natijalarni ko‘rsatadi. Safder va boshqalar oltmish xususiyat bilan o‘qitilgan chuqur neyron tarmog‘i bilan 93.32 foizlik F1 natijasiga erishgan. Bu ko‘rsatkich qoidalarga asoslangan modeldan yigirma sakkiz foiz va SVM modelidan o‘n olti foiz yuqoridir. Ushbu natijalar chuqur o‘rganish, xususan transformer modellarining metama’lumotlarni ajratishda eng istiqbolli yondashuv ekanligini ko‘rsatadi.

ISSN 3030-3907

Mavjud ochiq kodli tizimlarning qiyosiy tahlili shuni ko‘rsatadiki, GROBID metama’lumotlar va adabiyotlarni ajratishda eng yaxshi umumiy natijalarni qayd etgan, CERMINE esa o‘rtacha 77.5 foizlik F1 bilan undan keyin joylashgan. Anystyle bibliografik havolalarni ajratishda eng kuchli vosita sifatida namoyon bo‘lgan. Jadvallarni ajratish bo‘yicha Adobe Extract boshqalardan ustun bo‘lsa-da, uning umumiy samaradorligi pastligicha qolmoqda. Ro‘yxatlar, kolontitullar va tenglamalarni ajratishda esa barcha vositalar jiddiy qiyinchiliklarga duch keladi.

XULOSA.

Ushbu maqolada akademik maqolalardan metama’lumotlarni avtomatik ekstraksiya qilish muammosining nazariy asoslari, mavjud yechimlar va ularning qiyosiy tahlili keltirildi. Tadqiqot shuni ko‘rsatdiki, metama’lumotlar ilmiy ma’lumotlarni indeksatsiyalash, qidirish va bibliometrik tahlil qilishda hal qiluvchi ahamiyatga ega. Biroq ularni qo‘lda kiritish ko‘p vaqt talab etadi va xatolarga yo‘l qo‘yadi.

PDF formatining keng tarqalganligi, unda strukturaviy ma’lumotning yo‘qligi, skanerlangan hujjatlar muammosi va turli nashriyotlar shablonlaridagi xilma-xillilik metama’lumotlarni avtomatik ajratishni qiyinlashtiradi. LaTeX va XML formatlari mantiqiy strukturani aniq belgilaydi, ammo ular universal emas.

Mavjud yechimlar evolyutsiyasi qoidalarga asoslangan tizimlardan an’anaviy mashinaviy o‘rganish va so‘nggi yillarda chuqur o‘rganishga qarab rivojlanmoqda. Qoidalarga asoslangan tizimlar ma’lum jurnal formatlarida yuqori aniqlikka erisha oladi, ammo moslashuvchanligi past. An’anaviy mashinaviy o‘rganish modellari nisbatan universal, lekin katta hajmdagi anotatsiyalangan o‘quv ma’lumotlariga muhtoj. Chuqur o‘rganish modellari kontekstual va semantik tushunish qobiliyati tufayli murakkab va nostandart hollarda eng yaxshi natijalarni ko‘rsatadi.

Ochiq kodli vositalarning qiyosiy tahlili GROBID, CERMINE, Science Parse va Anystyle kabi tizimlar mavjudligini ko‘rsatsa-da, ularning hech biri barcha formatlar va tillar uchun universal va barqaror natija bermaydi. Skanerlangan hujjatlar, ko‘p tillilik - xususan o‘zbek tili - katta til modellarining yetarli darajada moslashtirilmaganligi va yuqori sifatli anotatsiyalangan korpuslarning kamligi sohadagi asosiy bo‘shliqlar hisoblanadi.

Kelgusidagi tadqiqotlar o‘zbek tilidagi ilmiy maqolalar uchun metama’lumotlarni ajratish tizimini ishlab chiqish, GPT-4 va BERT kabi katta til modellarini PDF ilmiy maqolalardan metama’lumotlarni ajratish vazifasiga moslashtirish, turli formatlar va tillarni qamrab oluvchi ochiq anotatsiyalangan korpus yaratish hamda skanerlangan hujjatlar uchun OCR va strukturaviy tahlilni birlashtiruvchi gibrid tizimlarni ishlab chiqishga qaratilishi kerak. Ushbu maqolada

ISSN 3030-3907

taqdim etilgan tahlil keyingi tadqiqotlar uchun mustahkam nazariy va amaliy asos bo‘lib xizmat qiladi.

FOYDALANILGAN ADABIYOTLAR

1. Rasuli, B., Boock, M., Schöpfel, J. et al. The link between dissertation metadata completeness and user engagement in an institutional repository. Scientometrics 130, 2875–2899 (2025). https://doi.org/10.1007/s11192-025-05331-0
2. T. J. Skluzacek, “Automated metadata extraction can make data swamps more navigable,” Ph.D. dissertation, The University of Chicago, 2022
3. T. J. Skluzacek, K. Chard and I. Foster, "Automated metadata extraction: challenges and opportunities," 2022 IEEE 18th International Conference on e-Science (e-Science), Salt Lake City, UT, USA, 2022, pp. 495-500, doi: 10.1109/eScience55777.2022.00088.
4. 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.
5. 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.
6. Manzoor, J.A.: Segmenting electronic theses and dissertations by chapters. MS thesis, Virginia Tech, Computer Science, defended September 23, 2022 (2022).
7. A. M. Hashmi, M. T. Afzal and S. u. Rehman, "Rule Based Approach to Extract Metadata from Scientific PDF Documents," 2020 5th International Conference on Innovative Technologies in Intelligent Systems and Industrial Applications (CITISIA), Sydney, Australia, 2020, pp. 1-4, doi: 10.1109/CITISIA50690.2020.9371784.
9. E. Beydaghi, M. Rahnama and J. A. Nasiri, "Ensemble approach for metadata extraction in Persian theses," 2020 6th Iranian Conference on Signal Processing and Intelligent Systems (ICSPIS), Mashhad, Iran, 2020, pp. 1-5, doi: 10.1109/ICSPIS51611.2020.9349571.

ISSN 3030-3907