TOSHKENT AXBOROT TEXNOLOGIYALARI UNIVERSITETI NUKUS FILIALI HUZURIDAGI ILMIY DARAJALAR BERUVCHI PhD.13/05.05.2023.T.162.01. RAQAMLI ILMIY KENGASH

MUHAMMAD AL-XORAZMIY NOMIDAGI TOSHKENT AXBOROT TEXNOLOGIYALARI UNIVERSITETI NUKUS FILIALI

KENJAYEV XAMDAM BAZARBAYEVICH

KALITLI KOMPONENTLARDAN FOYDALANGAN HOLDA BIR JINSLI HUJJATLARDAN MA'LUMOTLARNI AJRATIB OLISHNING ALGORITMIK VA DASTURIY MAJMUASI

05.01.04 – Hisoblash mashinalari, majmualari va kompyuter tarmoqlarining matematik va dasturiy ta'minoti

TEXNIKA FANLARI BO'YICHA FALSAFA DOKTORI (PhD) DISSERTATSIYASI AVTOREFERATI

Nukus-2024

UDK: 004.42:004.91:004.62

Texnika fanlari bo‘yicha falsafa doktori (PhD) dissertatsiyasi avtoreferati mundarijasi

Оглавление автореферата диссертации доктора философии (PhD) по техническим наукам

Contents of dissertation abstract of the doctor of philosophy (PhD) on technical sciences

MUHAMMAD AL-XORAZMIY NOMIDAGI TOSHKENT AXBOROT TEXNOLOGIYALARI UNIVERSITETI
NUKUS FILIALI HUZURIDAGI ILMIY DARAJALAR BERUVCHI PhD.13/05.05.2023.T.162.01. RAQAMLI ILMIY KENGASH

TOSHKENT AXBOROT TEXNOLOGIYALARI UNIVERSITETI NUKUS FILIALI

KENJAYEV XAMDAM BAZARBAYEVICH

KALITLI KOMPONENTLARDAN FOYDALANGAN HOLDA BIR JINSLI HUJJATLARDAN MA'LUMOTLARNI AJRATIB OLISHNING ALGORITMIK VA DASTURIY MAJMUASI

05.01.04 – Hisoblash mashinalari, majmualari va kompyuter tarmoqlarining matematik va dasturiy ta'minoti

TEXNIKA FANLARI BO'YICHA FALSAFA DOKTORI (PhD) DISSERTATSIYASI AVTOREFERATI

Nukus-2024

Texnika fanlari bo‘yicha falsafa doktori (PhD) dissertatsiyasi mavzusi O‘zbekiston Respublikasi Vazirlar Mahkamasi huzuridagi Oliy attestatsiya komissiyasida V2023.2.PhD/T.3636 raqam bilan ro‘yxatga olingan.

Dissertatsiya Toshkent axborot texnologiyalari universiteti Nukus filialida bajarilgan.

Dissertatsiya avtoreferati uch tilda (o‘zbek, rus, ingliz (rezyume)) Ilmiy kengash veb-sahifasida (www.tatunf.uz) va "Ziyonet" Axborot ta'lim portalida (www.ziyonet.uz) joylashtirilgan.

Ilmiy rahbar:	**Nishanov Axram Xasanovich** texnika fanlari doktori, professor
Rasmiy opponentlar:	**Djumanov Jamoljon Xudayqulovich** texnika fanlari doktori, professor
	Raximboev Xikmat Jumanazarovich texnika fanlari falsafa doktori (PhD), dotsent
Yetakchi tashkilot:	Buxoro davlat universiteti

Dissertatsiya himoyasi Toshkent axborot texnologiyalari universiteti Nukus filiali huzuridagi PhD.13/05.05.2023.T.162.01. Ilmiy kengashning 2024-yil "26" iyul soat 14⁰⁰ dagi majlisida bo‘lib o‘tadi. (Manzil: 230100, Nukus shahri, A.Dosnazarov ko‘chasi, 74-uy. Tel.: (861) 222-49-10,e-mail: tatunf@tatunf.uz).

Dissertatsiya bilan Toshkent axborot texnologiyalari universiteti Nukus filiali Axborot-resurs markazida tanishish mumkin (1 raqam bilan ro‘yxatga olingan.). (Manzil: 230100, Nukus shahri, A.Dosnazarov ko‘chasi, 74-uy. Tel.: (861) 222-49-10).

Dissertatsiya avtoreferati 2024-yil "15" iyul da tarqatildi.
(2024-yil "12" iyul dagi 1 raqamli reestr bayonnomasi.)

B.T.Kaipbergenov
Ilmiy darajalar beruvchi Ilmiy kengash raisi, texnika fanlar doktori, professor

R.I.Oteniyazov
Ilmiy darajalar beruvchi Ilmiy kengash ilmiy kotibi, texnika fanlar doktori, professor

K.K.Seitnazarov
Ilmiy darajalar beruvchi Ilmiy kengash huzuridagi ilmiy seminar raisi, texnika fanlar doktori, professor

KIRISH (falsafa doktori (PhD) dissertatsiyasining annotatsiyasi)

Dissertatsiya mavzusining dolzarbligi va zarurati. Jahonda AKT va Internet tarmog‘ining rivojlanishi natijasida insonlar extiyojini qanoatlandiruvchi elektron shaklga o‘tkizilayotgan qog‘oz shaklidagi axborot man’balri, xususan, hujjatlar, gazeta va jurnal kabi nashrlar o‘z o‘rnida katta hajm va oqimdagi axborotlarni boshqarish, qayta ishlash va foydalanuvchiga taqdim etish masalalarga alohida ahamiyat berilmoqda. Hozirgi kunda rivojlangan mamlakatlarda, jumladan, IT sohasida ilg‘or bo‘lgan AQSh, Angliya, Janubiy Koreya, Germaniya, Yaponiya, Rossiya kabi davlatlarning yetakchi kompaniyalar va olimlari olib borayotgan tadqiqotlarida tabiiy tildagi matnlarni sun’iy intellekt, mashinali o‘qitish va katta ma’lumotlarni tahlil qilish vositalari yordamida qayta ishlashning zamonaviy yechimlarini amaliyotga joriy qilish belgilangan. Bu borada, jumladan, kalit so‘zlar orqali bir jinsli to‘plam hujjatlarining matnlarini tabiiy tilga bog‘liq qayta ishlash, aniq ma’lumotlarni ajratib olish va hujjatlarni umumlashtirish uchun mo‘ljallangan maxsus tizim, dasturiy vositalar va algoritmlari tabora takomillashtirishga alohida e’tibor qaratilmoqda.

Jahonda katta auditoriyali olimlar tomonidan o‘xshash format yoki tuzilishga ega bo‘lgan hujjatlar tarkibidan ulardagi alohida elementlarni aniqlash va ajratib olish jarayoni ustida juda muhim ilmiy tadqiqotlar olib borilmoqda. Ushbu yo’nalishda, jumladan, bir xilligi format va tuzilmadagi hujjatlar bazasidan ularning qiymatli axborotlarini umumlashtirish va katta hujjatning mazmunini ma’nosini shakllantiruvchi algorotmik-matematik ta’minotlarini ishlab chiqish bo’yicha tadqiqotlar ustuvor hisoblanmoqda. Bunga ularni yagona formatga aylantirish va ahamiyatsiz ma’lumot-larni o‘chirish yoki xatolarni tuzatishlar dolzarb vazifalardan hisoblanmoqda.

Respublikamizda hujjatdagi tegishli bo‘limning boshlanishini ko‘rsatadigan maxsus iboralar yoki formatlarni izlash, tabiiy tilda berilgan matnlarni qayta ishlash, murakkab hujjatlarda kontekstni tushunish, ma’no chiqarish va matnning turli bo‘limlari o‘rtasidagi munosabatlarni aniqlash, algoritmlar, ayniqsa murakkab yoki turli xil ma’lumotlar to‘plamlarida, vaqt o‘tishi bilan ularning aniqligini oshirish uchun mashinali o‘qitishdan foydalanish, har bir ekstraksiya jarayonidan o‘rganashlar va tegishli ma’lumotlarni aniqlashlar, olingandan so‘ng, ma’lumotlarni tahlil qilish yoki boshqa tizimlarga, masalan, ma’lumotlar bazalari yoki ma’lumotlarni tahlil qilish vositalariga integratsiya qilish uchun mos formatga o‘tkazish va ekstraksiya tizimlari haqiqatan ham samarali bo‘lishi uchun mavjud ma’lumotlar bazalari, kontentni boshqarish tizimlari yoki biznes ish oqimlari bilan integratsiya qilinishi borasida yetarlicha ilmiy tadqiqotlar va isloxotlar olib borilmoqda. Xususan, O‘zbekiston Respublikasi davlat boshqaruv organlarida hujjatlar oqimini raqamli formatda boshqarish va tizimli qayta ishlash sohasida “…elektron hujjat almashinuvi … operatsion jarayonlarni raqamlashtirish...”[1], “…ta’lim jarayonini avtomatlashtirish bo‘yicha axborot tizimi … elektron

[1] O‘zbekiston Respublikasi Prezidentining 2020-yil 5-oktabrdagi ““Raqamli O‘zbekiston - 2030” strategiyasini tasdiqlash va uni samarali amalga oshirish chora-tadbirlari to‘g‘risida” PF-6079-son Farmoni.

shakllar…”, “… Xalq ta’limi vazirligi va uning hududiy bo‘linmalari faoliyatining boshqaruv jarayonlarini avtomatlashtirish…”[2] kabi vazifalar belgilangan. Ushbu vazifalarni amalga oshirishda o‘zbek tilidagi matnli hujjatlarni qayta ishlash, xususan bir xil andozali hujjatlarni mashina yordamida umumlashtirish va avtomatik hisobotlarni tayyorlash borasida ilmiy-amaliy tadqiqotlar olib borish va dasturiy majmualarini ishlab chiqsh muhim masalalardan hisoblanadi.

O‘zbekiston Respublikasining “Elektron hukumat to‘g‘risida”gi (2015) va “Jismoniy va yurdidik shaxslarning murojaatlar to‘g‘risida”gi (2017) Qonunlari, O‘zbekiston Respublikasi Prezidentining 2020-yil 5-oktabrdagi PF-6079-son “Raqamli O‘zbekiston-2030 strategiyasini tasdiqlash va uni samarali amalga oshirish chora-tadbirlari to‘g‘risida”, 2017 yil 7 fevraldagi PF-4947-son “O‘zbekiston Respublikasini yanada rivojlantirish bo‘yicha harakatlar strategiyasi to‘g‘risida”, 2020 yil 29 oktabrdagi PF-6097-son “Ilm-fanni 2030 yilgacha rivojlantirish konsepsiyasini tasdiqlash to‘g‘risida”gi farmonlari, 2021-yil 17-fevraldagi PQ-4996-son “Sun’iy intellekt texnologiyalarini jadal joriy etish uchun shart-sharoitlar yaratish chora-tadbirlari to‘g‘risida” qarori, Vazirlar Mahkamasining 2018-yil 5-yanvardagi 7-son “Fuqarolarning o‘zini o‘zi boshqarish organlarida jismoniy va yuridik shaxslarning murojaatlari bilan ishlash tartibi to‘g‘risida namunaviy nizomni tasdiqlash haqida” va 2018-yil 7-maydagi 341-son “Davlat organlarida, davlat muassasalarida va davlat ishtirokidagi tashkilotlarda jismoniy va yuridik shaxslarning murojaatlari bilan ishlash tartibi to‘g‘risidagi namunaviy nizomni tasdiqlash haqida” qarorlari hamda mazkur faoliyatga tegishli boshqa me’yoriy-huquqiy hujjatlarda belgilan vazifalarni amalga oshirishga ushbu tadqiqot muyaan xizmat qiladi.

Tadqiqotning respublika fan va texnologiyalari rivojlani-shining ustuvor yo‘nalishlariga mosligi. Mazkur tadqiqot Respublika fan va texnologiyalari rivojlanishining IV. “Axborotlashtirish va axborot-kommunikatsiya texnologiyalarini rivojlantirish” ustuvor yo‘nalishi doirasida bajarilgan.

Muammoning o‘rganilganlik darajasi. Kalitli komponentlardan foydalangan holda bir jinsli hujjatlardan ma’lumotlarni ajratib olishning algoritmik va dasturiy majmuasi kabi masalalarni xal qilishda bir qator xorijiy olimlar hissa qo’shganlar, jumladan, A.O.Shigarov, I.V.Bichkov, S.O.Sheremeteva, A.D.Ustalov, E.V.Stojok, A.E.Xmelnov, E.Yu.Xrustapev, A.V.Solovev, T.G.Penkova, H.H. Chen, S.C.Tsai, R.Campos, V.Mangaravite, A.Pasquali, A.Jorge, C.Nunes, A.Jatowt, Z.Jingshenglar[3] va boshqalar.

Kalitli komponentlardan foydalangan holda bir jinsli hujjatlardan ma’lumotlarni ajratib olish usullarini ishlab chiqish va takomillashtirishga O‘zbekistonni taniqli olimlarining ilmiy ishlari bag‘ishlangan. Bulardan:

[2] O‘zbekiston Respublikasi Prezidentining 2019-yil 29-apreldagi “O‘zbekiston Respublikasi Xalq ta’limi tizimini 2030-yilgacha rivojlantirish konsepsiyasini tasdiqlash to‘g‘risida” PF-5712-son Farmon.

[3] Christopher Manning, u Stenford universiteti bilan aloqador va tabiiy tilni qayta ishlash sohasida ko‘p ish qilgan. Uning “Tilni statistik tabiiy qayta ishlash asoslari” kitobi juda mashhur hisoblanadi; Jurafsky & Martin, ularning ‘Nutq va tilni qayta ishlash” kitobi butun dunyo bo‘ylab tarqalgan; Andrew Ng, uning asosiy ishi chuqur o‘qitish va uni turli sohalarda qo‘llash bilan bog‘liq bo‘lsada, mashinali o‘qitish sohasidagi hissasi ma’lumot olish usullariga tangensial ta’sir ko‘rsatdi; Jacob Devlin, Sebastian Riedel, Yoshua Bengio, Geoffrey Hinton, and Yann Le Chun, hozirda ular matnlardan ma’lumot olishda keng qo‘llaniladigan chuqur o‘qitish bo‘yicha tanilgan olimlardan hisoblanishadi.

M.M.Kamilov, Sh.X.Fazilov, R.H.Hamdamov, N.S.Mamatov, S.S.Radjabov, M.X.Xudoyberdiev, A.Hamroev, A.X.Nishanov va boshqalar.

Olib borilgan ilmiy tadqiqotlar natijasida kalitli komponentlardan foydalangan holda bir jinsli hujjatlardan ma'lumotlarni ajratib olishning algoritmik va dasturiy majmuasini amaliyotda qo‘llash masalalarini yechishda salmoqli natijalarga erishildi. Shu bilan birga, kalitli komponentlardan foydalangan holda bir jinsli hujjatlardan ma'lumotlarni ajratib olishning algoritmik va dasturiy majmuasi yo‘nalishining nazariy va amaliy muammolari yetarli darajada o‘rganilmagan.

Tadqiqotning dissertatsiya bajarilgan oliy ta'lim yoki ilmiy-tadqiqot muassasasining ilmiy-tadqiqot ishlari rejalari bilan bog‘liqligi. Dissertatsiya tadqiqoti Muhammad al-Xorazmiy nomidagi Toshkent axborot texnologiyalari universitetining ilmiy-tadqiqot ishlari rejasiga muvofiq №22/19-F “Milliy elektron hukumat muhiti xizmatlaridan foydalanish samaradorligini oshirish “iGov-maslahat-muhokama-monitoring tizimi”ni ishlab chiqish va uning tadbiqi” (2019-2020) va Kamoliddin Behzod nomidagi milliy rassomlik va dizayn institutining ilmiy-tadqiqot ishlari rejasiga muvofiq №333-U “O‘zbekiston badiiy ta'lim tizimida “Onlayn muzey” elektron galeriyasini yaratish” (2021-2022) hamda Muhammad al-Xorazmiy nomidagi TATU Nukus filialining ilmiy-tadqiqot ishlari rejasiga muvofiq №IL-392103072 “Chorvachilik komplekslarini elektron boshqarishning mobil ilovasini yaratish” (2022-2023) loyihalari doirasida bajarilgan.

Tadqiqotning maqsadi. Bir xil tuzulmaga ega matnli hujjatlardan kalit so‘zlar orqali muhim axborot birliklarini ajratib olishning va bu ma'lumotlardan ehtiyojga ko‘ra umumlashtiruvchi hisobotlar shakllantirishning algoritmik va dasturiy majmuasini ishlab chiqishdan iborat.

Tadqiqotning vazifalari:

kalit so‘zlar asosida hujjat axborat birliklarini umumlashtirish masalaning matematik modeli va dasturiy majmua infratuzilmasini ishlab chiqish;

elektron hujjat andozalari, kalit so‘zlar va bilimlar bazasini loyihalash;

elektron hujjatlar andozalari, fragmentlari, kalit so‘zlar asosida qiymatli axborotlarni ajratib olish va umumlashtitirish uslubiyati va algoritmik ta'minotini ishlab chiqish;

ishlab chiqilgan algoritmlar kesimida amaliy masalalarni yechishga ko‘mak beruvchi dasturiy majmuasi ishlab chiqish.

Tadqiqotning obyekti sifatida katta hajmdagi axborot resurslari, jumladan matnli hujjatlar to‘plamidan foydalanuvchi ehtiyojini qanoatlandiruvchi zarur axborotlarni qidirish, tahlil qilish va ularni umumlashtirishga asoslangan yondashuvlar olingan.

Tadqiqotning predmeti elektron hujjat aylanish tizimlaridan chiquvchi so‘rov hujjatga kelgan ko‘p sonli hujjatlarni umumlashtirish yoki aralash matnli hujjatlar to‘plamidan belgilangan andoza va kalit so‘zlarga muvofiq zarur axborot birliklarini chiqarib olish algoritmlari va qaror qabul qilishga ko‘maklashuvchi dasturiy majmuadan iborat.

Tadqiqotning usullari. Tadqiqot jarayonida ma’lumotlarni intellektual tahlil qilish, katta hajmli ma’lumotlarga ishlov berish va timsollarni tanib olish nazariyasi usullaridan foydalanilgan.

Tadqiqotning ilmiy yangiligi quyidagilardan iborat:

elektron hujjatning fragment andozalari, kalit so‘zlar va bilimlar bazasi asosida matnlarni qayta ishlash usullari loyhalashtirilgan;

hujjatlardan toifali fragment bloklarini aniqlash va bu bloklardan kalit so‘zlar asosida qiymatli axborotlarni ajratib olish jarayonining matematik modeli ishlab chiqilgan;

o‘zbek tilida ifodalangan matnli hujjatlarni fragmentlash, fragment matnidan kalit so‘zlar asosida qiymatli axborotlarni ajratib olish va umumlashtirishning qoidalar bazasi ishlab chiqilgan;

bir jinsli elektron hujjatlarning fragment matnidan qiymatli axborotlarni ajratib olish va umumlashtirish qoidalar bazasi asosida qiymatli axborotlarni chiqarib olish va umumlashtirish algoritmlari ishlab chiqilgan.

Tadqiqotning amaliy natijalari quyidagilardan iborat:

katta hajmdagi hujjatlar bilan shug‘ullanadigan sohalarda hujjatlarni umumlashtirish, qaror qabul qilish, tahliliy hisobot jarayonlarni yaxshilash mumkinligi asoslangan;

avtomatlashtirilgan tizim qo‘lda ishlov berishga qaraganda katta hajmdagi hujjatlarni boshqarishi, ma’lumotlarni muntazam va tez qayta ishlovchi tashkilotlar ish samaradorligini oshirish uchun ishlab chiqilgan;

hujjatlardan asosiy tarkibiy qismlarni samarali ajratib olish orqali tegishli ma’lumotlar yanada qulayroq bo‘lib, bilimlarni boshqarish va ma’lumotlarni qidirish jarayonlarini qo‘llab-quvvatlashi mumkinligi asoslangan;

bunday tizimlar muayyan sohalar yoki hujjatlar turlariga moslashtirilishi mumkin, bu ularni turli ilovalar uchun ko‘p qirrali vositalarga aylantiradi. Masalan, ular sog‘liqni saqlash sohasidagi bemor ma’lumotlarini yoki bankdagi tranzaksiya tafsilotlarini olish uchun moslashtirilishi mumkinligi asoslangan;

Xulosa qilib aytadigan bo‘lsak, bir xil tipdagi hujjatlardan ma’lumot olish uchun algoritmik va dasturiy majmua turli sektorlarda ma’lumotlarni qayta ishlash samaradorligini, aniqligini va tezligini oshirish, shu bilan birga turli ehtiyojlarni qondirish uchun masshtablash va moslashtirish imkoniyatlarini taklif qilish qobiliyati tufayli juda katta amaliy ahamiyatga ega.

Tadqiqot natijalarining ishonchliligi. Tadqiqot yakunida o‘z aksini topgan umumnazariy xulosalar, o‘xshash tuzilishdagi hujjatlar to‘plamidan tegishli ma’lumotlarni avtomatik ravishda ajratib oladigan tizimning ishlash aniqligi, ma’lumotlarni yig‘ish va qayta ishlash, bir xil tipli hujjat ma’lumotlarini umumlashtirish, ma’lumotlarni qayta ishlash, algoritm loyihalash ishlab chiqish, algoritmlarni sinab ko‘rish, dasturiy ta’minotni kodlashning eng yaxshi amaliyotlari, modullilik va masshtablilik, tekshirish, testlash, xatolarni tahlil qilish va takomillashtirish kabi faktorlar bilan izohlanadi.

Tadqiqot natijalarining ilmiy va amaliy ahamiyati.

Tadqiqot natijalarining ilmiy ahamiyati o‘xshash hujjatlar to‘plamidan tabiiy tildagi matnlardan asosiy komponentlarni aniqlash va tahlil qilish orqali tegishli ma’lumotlarni chiqarib olish, mashinali o‘qitish usullarini takomillashtirish, taqqoslash, bir xil hujjatlar turlarni umumlashtirish samaradorligi va aniqlik mezonlarni bilan izohlanadi.

Tadqiqot natijalarining amaliy ahamiyati ma’lumotlarni olish jarayonlarini avtomatlashtirish, axborotni qidirishning aniqligini oshirish, turli sohalarda qo‘llanilishi, bilimlarni boshqarish tizimlarini takomillashtirish, kontent tahlilini osonlashtirish, katta miqyosdagi ma’lumotlarni tahlil qilishni yoqish, masshtablilik va samaradorlik, moslashuvchanlik, hisoblash tilshunosligi sohasida hissasi va soha mutaxassislariga zamonaviy ko‘mak beruvchi dasturiy tizim ishlab chiqilganligi bilan izohlanadi.

Tadqiqot natijalarining joriy qilinishi. Ilmiy tadqiqotda ishlab chiqilgan algoritmlar asosida yaratilgan “ARS-Uz” dasturiy majmua asosida:

o‘zbek tilida ifodalangan matnli hujjatlarini fragmentlash, fragment matnidan kalit so‘zlar asosida qiymatli axborotlarni ajratib olish va umumlashtirishning qoidalar bazasi asosida ishlab chiqilgan axborot tizim Ellikqal’a tumani Maktabgacha va maktab ta’lim bo‘limiga joriy etildi (Qoraqalpog‘iston Respublikasi Maktabgacha va maktab ta’limi vazirligining 2023 yil 02 noyabrdagi №01-03/4034-son ma’lumotnomasi). Natijada tashkilotning yuqori hamda quyi bo‘g‘inidagi tashkilotlaridan keladigan turli ko‘rinishdagi axborotlar (xat, buyruq va hokazolar) hamda fuqoralar murojaatlarini tahlil qilish va tezkorlikda javob qaytarish kabi jarayonlar uchun ketadigan vaqtni 21% ga qisqartirgan. Bu esa ish samaradorligini 23% ga oshirish imkonini bergan;

elektron hujjat tuzilishi va fragment andozalari, kalit so‘zlar va bilimlar bazasini loyhalashtirish va hujjatlardan toifali fragment bloklarini aniqlash hamda bu blokdan kalit so‘zlar asosida qiymatli axborotlarni ajratib olish jarayoni matematik modeli qurish asosida yaratilgan dasturiy majmuasi Beruniy tumani Maktabgacha va maktab ta’lim bo‘limiga joriy etildi (Qoraqalpog‘iston Respublikasi Maktabgacha va maktab ta’limi vazirligining 2023 yil 02 noyabrdagi №01-03/4034-son ma’lumotnomasi). Natijada tashkilotga keladigan elektron hujjatlardan toifali fragment bloklarini aniqlash va bu blokdan kalit so‘zlar asosida qiymatli axborotlarni ajratib olish jarayoni avtomatlashtirilgan holda ishlash imkonini bergan. Bu esa soha xodimlarini ish yuklama hajmini 17% ga qisqartirgan hamda keladigan hujjatlarga qayta ishlov berish asosida javob qaytarish tezkorligini oshirish orqali ish samaradorligini 19% ga oshirish imkonini bergan;

elektron hujjatlarni kalit so‘zlar asosida qiymatli axborotlarni ajratib olish va umumlashtirish algoritmi asosida yaratilgan dasturiy majmuasi Amudaryo tumani Maktabgacha va maktab ta’lim bo‘limiga joriy etildi (Qoraqalpog‘iston Respublikasi Maktabgacha va maktab ta’limi vazirligining 2023 yil 02 noyabrdagi №01-03/4034-son ma’lumotnomasi). Natijasida tashkilot xodimlari uchun 7 xil kategoriyaga tegishli kalit so‘zlar bazasi asosida hujjatlarni qabul qilib olish va avtomatik tahlil qilish jarayonnini bajarish hamda talab asosida har bir hujjatdagi

qiymatli ma’lumotlarni ajratib berish kabi imkoniyatlarni yaratgan. Bu esa vaqtni 22% ga qisqartirish orqali ish samaradorligini 24% ga oshirish imkonini bergan.

Tadqiqot natijalarining aprobatsiyasi. Dissertatsiyaning asosiy nazariy hamda amaliy natijalari 4 ta xalqaro va 7 ta respublika ilmiy–texnik hamda ilmiy–amaliy anjumanlarida muhokama qilingan.

Tadqiqot natijalarining e’lon qilinganligi. Tadqiqot mavzusi bo‘yicha asosiy natijalar 23 ta ilmiy ishlarda e’lon qilingan bo‘lib, ulardan 8 ta O‘zbekiston Respublikasi Oliy attestatsiya komissiyasining doktorlik dissertatsiyalarining asosiy ilmiy natijalarini chop etish tavsiya etilgan ilmiy nashrlarda, jumladan 14 tasi respublika va 6 tasi xorijiy jurnallarda hamda 3 ta EHM uchun yaratilgan dasturiy mahsulotlarga qayd qilish guvohnomalari O‘zbekiston Respublikasi Adliya vazirligi huzuridagi intellektual mulk Agentligidan olingan.

Dissertatsiyaning tuzilishi va hajmi. Dissertatsiya kirish qismi, to‘rtta bob, xulosa, foydalanilgan adabiyotlar ro‘yxati va ilovalardan iborat. Dissertatsiyaning hajmi 110 betni tashkil etadi.

DISSERTATSIYANING ASOSIY MAZMUNI

Dissertatsiyaning kirish qismida tadqiqot mavzusining dolzarbligi, muammoning o‘rganilganlik darajasi, tadqiqotning maqsadi, vazifalari, obyekti, predmeti, tadqiqot metodlari va ilmiy farazi, himoyaga olib chiqilayotgan asosiy holatlar va ilmiy yangiligi, nazariy va amaliy ahamiyati, natijalarning joriy qilinishi va ishning sinovdan o‘tishi, natijalarning e’lon qilinganligi, foydalanilgan model va algoritmlar, qo‘llanilish sohasi va amalga oshirish bosqichlari qisqacha bayon etilgan.

Dissertatsiyaning birinchi bobi “**Hujjatlardan ma’lumotlarni ajratib olishda algoritmik yondashuvlar**” deb nomlangan bo‘lib, u hujjatlardan ma’lumotlarni ajratib olish, ya’ni, kalit so‘zlarni shakllantirish, matnlarni o‘zaro o‘xshashliklarini baholash algoritmlari va usullari tahlili keltirilgan. Shuningdek, matnli ma’lumotlardan obyektni nomini tanib olish yondashuvlari va hujjatlardan jadvallarni ajratib olish hamda matnlarni umumlashtirining avtomatlashgan usullari tahlili taqdim etilgan.

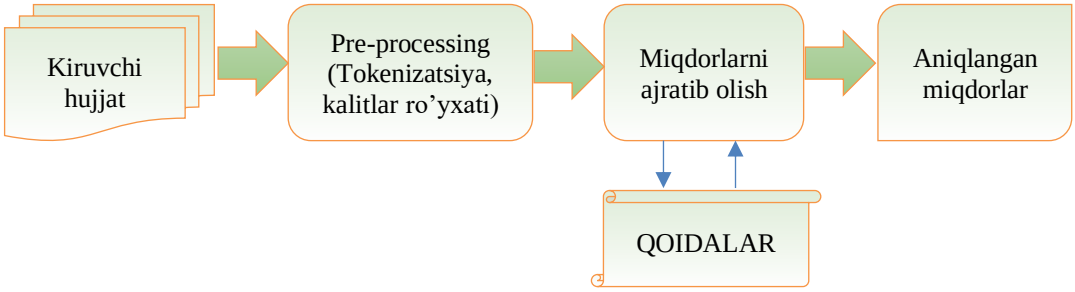
Hujjatlardan kalit so‘zlarni ajratib olish algoritmlari va matnlar o‘xshashligini baholash usullari tabiiy tilni qayta ishlash (NLP), tadqiqotlarda kalit so‘zlarni (KS) shakllantirish va ular asosida matnlarni umumlashtirish, matn tasnifi va klasterlash masalalarini hal etishda muhim ahamiyat kasb etadi. Bunda hujjatlardan kalit so‘zlarni ajratib olish algoritmlari matnlarning o‘zaro o‘xshashligini o‘lchash usullar tasnifi, inson va algoritm kalit so‘zlari o‘xshashligini baholash usullari tahlili keltirilgan.

Shuningdek, keyingi bandda matnli ma’lumotlrdan nomlangan obyektni tanib olish (NER) yondashuvi, qoidaga asoslangan nomlangan obyektni tanib olish (Named Entity Recognation-NER) asoslari va mavjud usullari qiyosiy tahlil qilingan va qoidaga asoslangan NER ning avzalliklari keltirib o‘tilgan. NER - bu ma’lum

ma’lumotlar to‘plami yoki korpusdan shaxs ismlari, tashkilotlar, vaqt, joylashuv kabi nomlangan obyektlarni aniqlash masalasi bilan shug‘ullanadi. Nomlangan obyektlar, predmet soha obyektlari (tibbiy, oziq-ovqat), korpusda belgilangan nomli obyektlar kabilarni o‘z ichiga oladi. Masalan:

Matn:
Xurshid 2023-yil Oksfordda o‘qish uchun 25000$ grant yutib oldi.
Chiqish:
Xurshid[Odam] 2023-yil[Vaqt] Oksfordda [Tashkilot] o‘qish uchun 25000$ [Miqdor son] grant yutib oldi.

NERning asosiy uchta usuli mavjud: lug‘atga asoslangan yondashuv, o‘qitishga asoslangan yondashuv (learning based) va qoidalarga asoslangan yondashuv (rule-based).



1-rasm. Qoidalarga asoslangan NER yordamida miqdorlarni aniqlash jarayoni.

Lug‘atga asoslangan NER hujjatlardan ma’lumot olish uchun ishlatiladi, chunki u tanib olingan termlar bo‘yicha ID ma’lumotlarini taqdim etishi mumkin. Ushbu usul termlarni moslashtirish orqali nomlangan obyektlarni aniqlaydi. Lug‘atga asoslangan yondashuvlar noto‘g‘ri ijobiy tanib olish va yangi nashr etilgan nomlarni qamrab oladigan yagona resursning yo‘qligi kabi cheklovlarga ega. Bu usul yuqori aniqlik darajasiga ega bo‘lsada, lekin u faqat lug‘atga kiritilgan NERni taniy oladi.

Ushbu tizimlar ba’zi cheklovlarga ega bo‘lsada, ular qimmat protativ emas va predmet sohaga bog‘liq. Bunda predmet soha bilimlari va dasturlash qobiliyatlari uchun inson tajribasi talab qilinadi. Qoidalarga asoslangan NER tizimlari faqat bitta predmet soha uchun mo‘ljallangan va boshqa predmet sohalarga ko‘chirilmaydi.

Qoidalarga asoslangan NERning afzallik jihati shundan iboratki, ma’lum soha uchun tabiiy tildagi har bir o‘zgachaliklarga ekspertlar qoidalar yaratishi mumkin.

NER yordamida hujjatdagi matnlardan nomlangan obyektlarni ajratib olish mumkin bo‘lsada, uni hujjatdagi jadvalli ma’lumotlarga qo‘llab bo‘lmaydi. Shu sababli hujjatdagi jadvallardan ma’lumotlarni ajratib olishning jadval modeli keying bo‘limda tadqiq qilinadi.

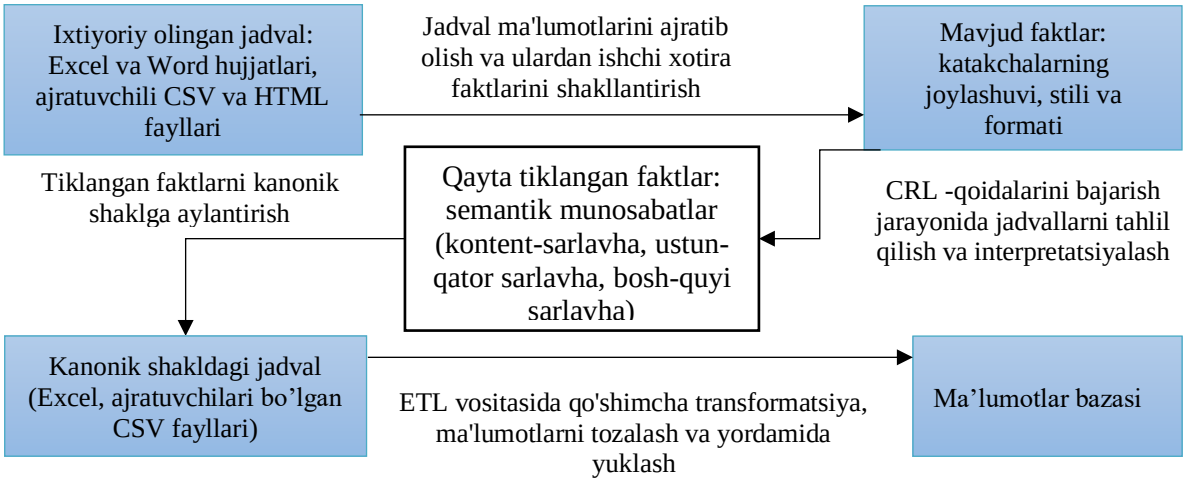
Taklif etilayotgan jadval modeli mantiqiy xulosa chiqarish jarayonida jadvallar haqidagi faktlarni taqdim etishga mo‘ljallangan. Model ikki darajadan iborat: jismoniy va mantiqiy.

Jismoniy daraja yacheykalarning geometrik pozitsiyalarini, uslublarini (grafik formatlash) va mazmunini tavsiflaydi. Bu daraja $T_p = (S_r; S_c; C)$ quyidagi

to‘plamlardan iborat: S_r qatorlar to‘plami va S_c - ustunlar to‘plami; $C - c = (c'; p; G)$ quyidagilarni o‘z ichiga olgan yacheykalar to‘plamidir: c' - kontent (qiymat); $p = (c_l; r_t; c_r; r_b)$ - S_r qatorlar va S_c ustunlaridagi koordinatalar (c_l - chap ustun, r_t - yuqori qator, c_r – o‘ng ustun va r_b - pastki qator); G - uslub sozlamalari to‘plami (shrift ko‘rsatkichlari, ranglar, matnni tekislash, chegara uslublar va boshqalar);

Mantiqiy daraja semantik munosabatlarni taqdim etadi (ya’ni, yacheyka- rol, sarlavha-qiymat, sarlavha-sarlavha va sarlavha-o‘lcham juftlari). Bu daraja $T_l = (D; L_r; L_c; E)$ quyidagi to‘plamlardan iborat: $D = \{D_i\}$ - qayta ishlangan jadvalda keltirilgan o‘lchamlar to‘plami. Ularning har biri $D_i = \{d_j\}$ o‘lchov qiymatlari to‘plami; L_r - qator sarlavhalar daraxti va L_c - ustun sarlavhalar daraxti. Bu daraxtlar ularning sarlavhalari orasidagi munosabatlarni taqdim etadi. Har bir sarlavha $l = (l')$ kontentga ega, l' bu D_i o‘lchovlarining qiymati emas: $l' \notin \bigcup D_i$. $E - e = (e'; D'; L')$ yozuvlar to‘plami: e' - mazmun; D' - bu yozuv bilan bog‘liq D_i o‘lchov qiymatlari to‘plami; L' - bu yozuv bilan bog‘liq L_r va L_c daraxtlaridan olingan sarlavhalar to‘plami.

Hujjatlardan jadvallarni olishning interpretatsiyalash bosqichida jadvallarni tahlil qilish va interpretatsiyalash uchun CRL (Cells Rule Language) deb nomlangan rasmiy qoidalar tili taklif qilingan. Bunda ixtiyoriy yarim strukturalangan jadvallardan ma’lumotlarni ajratib olish va ularni standart ETL (Extract, Transform, Load) vositalari yordamida ma’lumotlar bazasiga yuklash masalasi ko‘rib chiqilgan. Strukturalanmagan jadval ma’lumotlarini interpretatsiyalash uchun taklif etilayotgan jarayon sxemasi 2-rasmda ko‘rsatilgan.

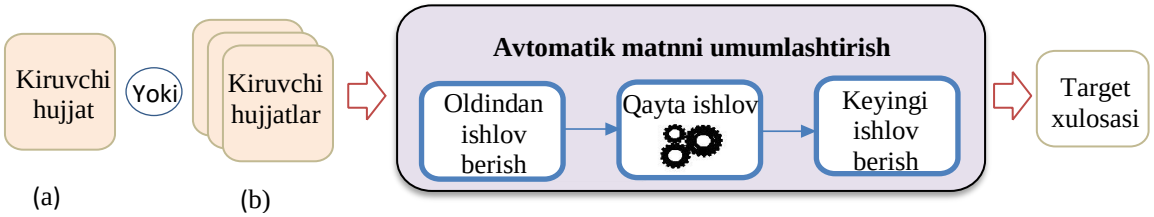


2-rasm. Jadvallarni tahlil qilish va interpretatsiyalash qoidalarini bajarish orqali strukturalanmagan jadval ma’lumotlarini interpretatsiyalash sxemasi.

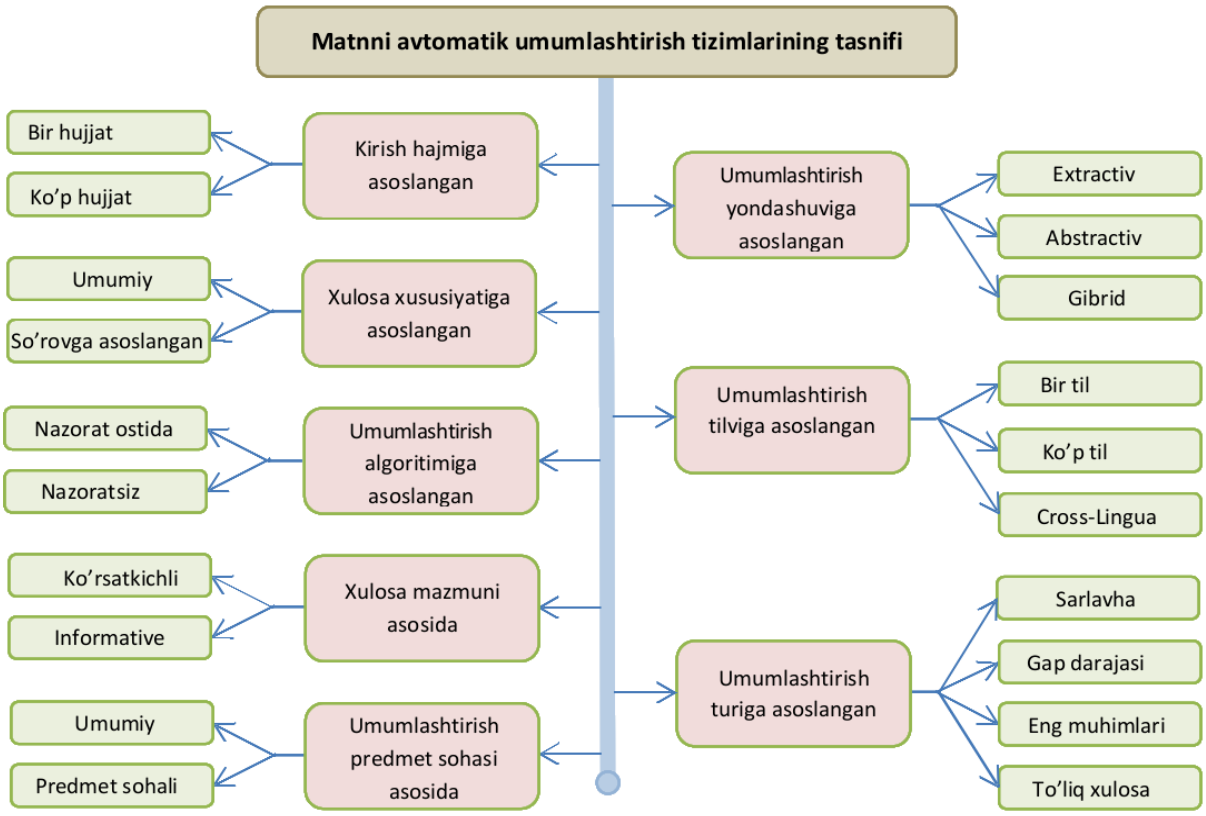
Jadvalni ajratib olish tuzilmagan ma’lumotlarni tuzilmali va amalda bo‘lgan formatga aylantirishda hal qiluvchi rol o‘ynaydi, ma’lumotlarni yanada qulayroq, tahlil qilinadigan va turli sohalar va turli ilovalar uchun muhim hisoblanadi. Hozirgi kunda hujjatlardan jadvallarni chiqarish uchun bir nechta dasturiy ta’minot vositalar amaliyotga joriy etilgan.

Dissertatsiya ishining ikkinchi bobi **“Bir jinsli hujjatlardan ma’lumotlarni ajratib olish usul va algoritmlari”** deb nomlangan bo‘lib, unda axborot tizimlari orqali hujjatlarni qayta ishlash yo‘nalishlari (dastur orqali hujjatni avtomat generatsiya qilish hamda hujjatlarni avtomatik tizimga o‘qitish) va ularda ma’lumotlarni taqdim etish usullari tadqiq etiladi. Jumladan, andozali hujjatlar, hujjat tarkibi va uni fragmentlash asoslari, hujjatlarni tizim ma’lumotlar bazasi yoki manba jadvalli asosida generatsiya qilish hamda tashqi manba elektron hujjatlarini mashina yordamida o‘qitish usul va algoritmlari batafsil qaraladi. Tadqiqot natijasida muammoni yechishga yo‘naltirilgan usullar taklif etiladi. Demak, mazkur ishda uchala masala (axborot taqdim etilish, avtomat generatsiya va hujjatlarni tizimga o‘qitish) bo‘yicha olib borilgan tadqiqot ishlari batafsil tahlil qilinadi va universal hujjatlar bilan ishlovchi tizimga talablar taklif etiladi.

Tadqiqotda asosiy masala bo‘lib matnli hujjatlar to‘plamidan tegishli axborot birliklarini chiqarib olish va yakuniy hujjatga ularni umumlashtirish masalasi qaraladi. Avtomatik matnni umumlashtirish (ATS) tizimining asosiy maqsadi - kirish hujjatining asosiy mazmunini o‘z ichiga olgan qisqa ma’lumotni kamroq joyda va takrorlanishni minimal darajada ushlab turishdir.



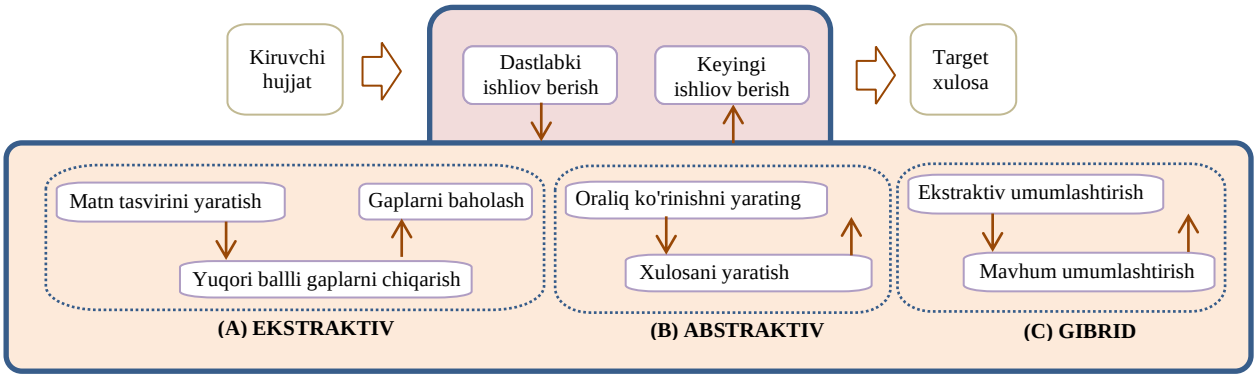
3-rasm. (a) bitta hujjatli yoki (b) ko‘p hujjatli avtomatik matn umumlashtiruvchi.



4-rasm. ATS tizimlarining tasnifi.

ATS tizimlarida ko‘plab tasniflar mavjud. Ushbu tasniflar haqida batafsil tizimli tahlil etilgan. ATS tizimlarini asosan quyidagi mezonlarga qarab tasniflash mumkin (4-rasm).

Umumlashtirish tasniflashning asosiysi uning yondashuvi (ekstraktiv, mavhum yoki gibrid) bo‘yicha bo‘ladi. ekstraktiv matnni umumlashtirish yondashuvi kirish hujjat(lar)idagi eng muhim gaplarni tanlaydi va bu tanlangan gaplar xulosada birlashtiriladi. Gibrid matnni umumlashtirish yondashuvi ekstraktiv va mavhum yondashuvlarning kombinatsiyasi (5- rasm).



5-rasm. Matnni umumlashtirish yondashuvlarining arxitekturasi.

ATS tizimlarini loyihalash va amalga oshirishda turli komponenta va usullar foydalaniladi.Umumiy holda hujjat quyidagicha tavsiflanishi mumkin:

$$f(\alpha, \beta), \qquad (1)$$

bunda α – hujjat parametrlari matritsasi, β – hujjat strukturasini tasvirlash matritsasi.

Hujjat $I \times k$ o‘lchamidagi α turli parametrlarning qiymatlar jadvali bo‘lib hisoblanadi:

$$\alpha = \begin{bmatrix} p_{1,1} & \cdots & p_{1,I} \\ \vdots & \ddots & \vdots \\ p_{k,1} & \cdots & p_{k,I} \end{bmatrix}, \qquad (2)$$

bu yerda p parametr qiymati.

Hujjatning $m \times n$ o‘lchamidagi β ma’lum tartibda tartiblangan hujjat fragmentlari jadvali bo‘lib, u quyidagicha ifodalanadi:

$$\beta = \begin{bmatrix} f_{1,1}(\Omega, \theta, \gamma) & \cdots & f_{1,m}(\Omega, \theta, \gamma) \\ \vdots & \ddots & \vdots \\ f_{n,1}(\Omega, \theta, \gamma) & \cdots & f_{n,m}(\Omega, \theta, \gamma) \end{bmatrix}, \qquad (3)$$

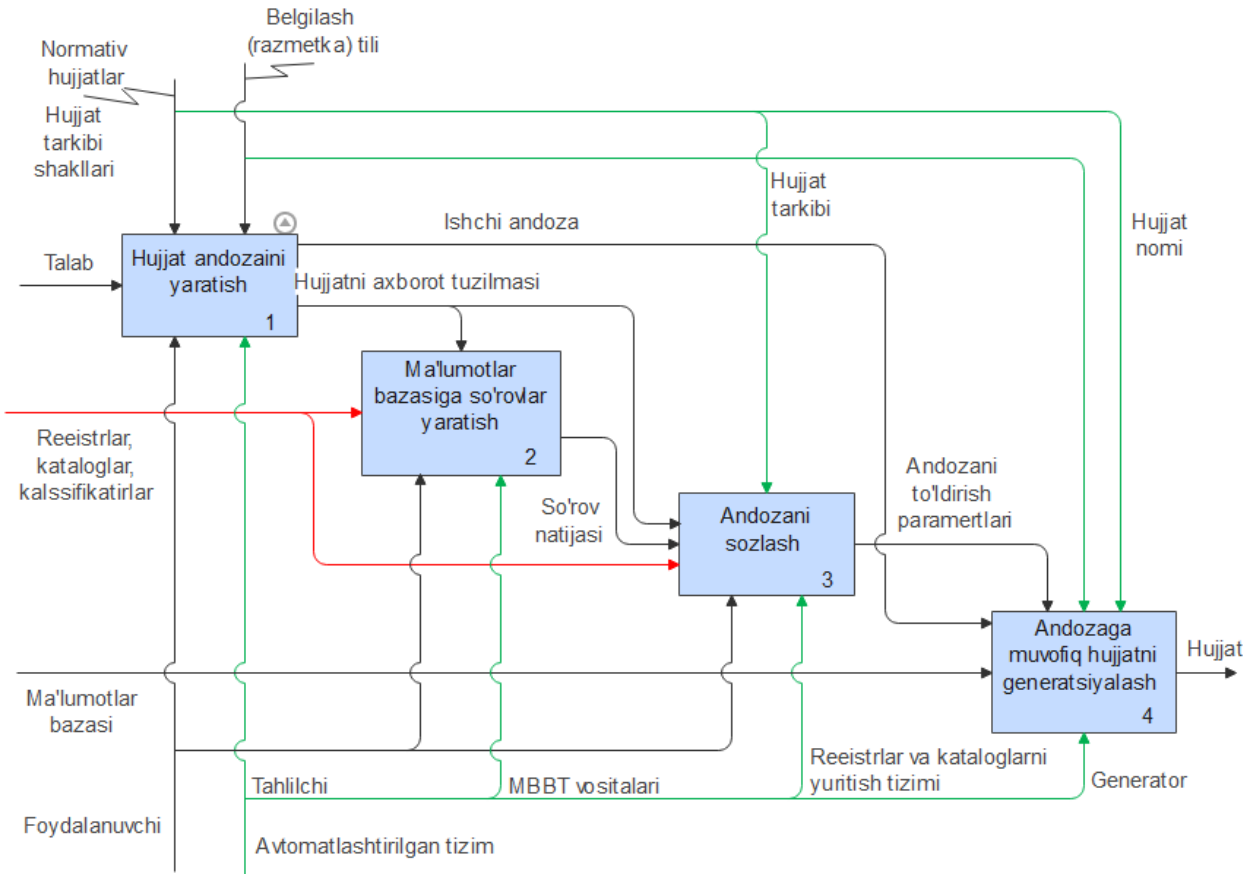
bunda f – hujjat fragmenti, Ω – ma’lumotlarni olish qoidasi, θ – fragment strukturasi (andoza), γ – qurish va vizualizatsiya qoidasi (algoritmi). α va β matritsalari Ω orqali o‘zaro bog‘liq bo‘lib, bunda α matritsasi qatorining ayrim elementlari Ω parametrlar to‘plamining elementlaridir. Bu berilgan ma’lumotlar massivini o‘zgartirish orqali hujjatning fragmentlarini o‘zgartirish imkonini beradi. Shunday qilib, Ω parametrlar to‘plami uchun matritsaning i-qatordagi elementlarining kichik to‘plami – M_i va ma’lumotlar massivini olish qoidasi – $\Omega(M_i)$ shaklida bo‘ladi, bunda $M_i \subset P_i$, $P_i \in \{p_{i,1} \dots p_{i,I}\}$.

Bir xil strukturali fragmentlar uchun ma’lumotlar massiviga qo‘shimcha ravishda qurilish va vizualizatsiya qoidalarning parametrlari o‘zgartirilishi mumkin. Ma’lumotlar massiviga o‘xshash $\phi(y)$ parametrlar funksiyasi kiritilsa, vizualizatsiya qoidasi quyidagicha bo‘ladi: $\gamma(\phi(y))$.

Endi hujjat strukturasini ifodalsh matritsasini (4) formula bilan berishladi,

$$\begin{bmatrix} f_{1,1}(\Omega_j(M_i),\theta_g,\gamma_k(\phi(y))) & \cdots & f_{1,m}(\Omega_j(M_i),\theta_g,\gamma_k(\phi(y))) \\ \vdots & \ddots & \vdots \\ f_{n,1}(\Omega_j(M_i),\theta_g,\gamma_k(\phi(y))) & \cdots & f_{n,m}(\Omega_j(M_i),\theta_g,\gamma_k(\phi(y))) \end{bmatrix}, \quad (4)$$

bunda $i \in I$ – hujjat parametrlar matritsasining qator raqami, $g \in G$ – fragmentni taqdim etish shaklining raqami, $j \in J$ – hujjat fragmentini qurish uchun berilgan ma’lumotlarini tanlash qoidasining raqami, $k \in K$ – hujjat fragmentini vizuallashtirish qoidasi raqami. Bu (4) formula texnologiyani ishlab chiqish, ko‘plab shakllarni saqlash va amalga oshirishda uning oldingi qaralgan talablarga muvofiqligini ta’minlaydi. Natijada, hujjatlarni formallashtirish dasturiy vositasi ishlash sxemasi taklif etilib, hujjatlarni o‘qish va taqdin etish texnologiyasi ishlab chiqiladi.



6-rasm. Hujjatlarni generatsiyalash jarayonining funksional modeli

Shuningdek, ushbu bobda hujjatlarni generatsiya qilish jarayonining funsional modeli ishlab chiqildi. Bu to‘rtta asosiy bosqichdan iborat: hujjat andozaini yaratish; ma’lumotlar bazasiga so‘rovlar yaratish; andozani to‘ldirish parametrlarini o‘rnatish va andozaga muvofiq hujjatni generatsiyalash. **Andozani yaratish** bosqichida foydalanuvchi tomonidan har bir hujjat turi uchun vizual maket yaratiladi,

ma’lumotlar bazasiga so‘rovlar yaratish bosqichda yaratilgan andozani to‘ldirish uchun ma’lumotlarni tayyorlanadi, **andozani sozlash** bosqichida andoza elementlari manba jadval ma’lumotlari yoki so‘rovlardan olingan ma’lumotlar bilan taqqoslanadi va **andozaga muvofiq hujjatni generatsiyalash** bosqichida kiritilgan maxsus maydonlar foydalanuvchi tomonidan ko‘rsatilgan parametrlar va shartlarga muvofiq avtomatik to‘ldiriladi.

Dissertatsiya ishining uchinchi bobi **“O‘zbek tilida ifodalangan hujjatlarni kalitli komponentalar asosida umumlashtirish”** deb nomlangan bo‘lib, unda tadqiq qilingan uslubiyatlarga, hususan kalit sozlar, NER, ATS va jadvallarni birlashtirish mexanizm va algoritmlarga tayangan holda o‘zbek tilida ifodalangan matnli elektron hujjatlarni kalit so‘zlar komponentalari asosida umumlashtirishning yangi uslubiyati, matematik asoslari bayon etiladi.

Hujjatni inson intellekti asosida umumlashtirish jarayoni misolidan ko‘rinib turibdiki, qaralayotgan masala uchun obyekt turlari va ularning nomlanishlari, predmet sohaga oid termlar (atamalar), ko‘rsatkichlarni ifodalovchi kalit so‘zlar bazasini (KSB) qurish zarur ekan. Bu KSB quyidagicha belgilanadi:

$$KSB = \{L, O, P, T, B, R, Q\}, \quad (5)$$

bu yerda:

$\boldsymbol{L} = \{l_1, l_2, \dots, l_{nl}\}$ – obyekt nomlanishlari to‘plami (Kegeyli, Nukus, …). Obyekt joylashuv manziliga ega jismoniy/virtual va mavjud tashkilot/hudud.

$\boldsymbol{O} = \{o_1, o_2, \dots, o_{no}\}$ – obyekt turlari to‘plami (tuman, mahalla, maktab …). Obyekt turi orqali hujjatdan obyekt nomi aniqlanadi;

$\boldsymbol{P} = \{p_1, p_2, p_3, \dots, p_{np}\}$ – predmet sohalar (qishloq xo‘jaligi, xalq ta’limi);

$\boldsymbol{T} = \{t_1, t_2, \dots, t_{nt}\}$ – tabiiy tilida termlar to‘plami (tadbirkor, ishsiz, qishloq, yer, kitob, o‘quvchi, fuqaro, davomat, nafaqa, suv, yoz,..);

$\boldsymbol{B} = \{b_1, b_2, \dots, b_{nb}\}$ – obyekt turlariga tegishli termlar (birikmasi) va uning sinonimlari to‘plami, buni obyekt kalit so‘zlari deyish mumkin (davomat (ishtirok etish, darsga kelish..), o‘quvchi (talaba, tolib, tinglovchi,…), o‘qituvchi (pedagog, muallim, domla, trener, …), ...);

$\boldsymbol{R} = \{r_1, r_2, \dots, r_{nr}\}$ – sonlar to‘plami (0,…, 9, bir, ikki, … o‘n, yuz, ming, million, milliard, I, V. X, C…).

$\boldsymbol{Q} = \{q_1, q_2, \dots, q_{nq}\}$ –qisqartmalar to‘plami (TATU, QDU, UzMU MTM, …)

Bu belgilanishlar KSB deb nomlanadi. KSB orqali dastlab obyekt va uning nomi aniqlanadi, keyin tanlangan predmet sohaga tegishli termlar asosida axborot birliklari chiqarib olinadi.

Hujjatdan fragmentlarni ajratib olish quyidagi funksiya orqali bajariladi:

$$\Theta\left(H_h, A_\varphi, \Psi_\varphi\right) \to X_i^{(j)}, \quad (6)$$

bunda H_h – kiruvchi to‘plamning h-hujjati, A – tanlangan andoza, Ψ – fragmentlash qoidasi, $X_i^{(j)}$ – ajratilgan (j)-fragment bloki, i-ushbu fragment turining uchrashi.

Umumiy holda axborot birliklarini chiqarib olish funksiyasi quyidagicha ifodalanadi:

$$F\left(X_i^{(j)}, \Gamma^{(j)}, FKS\right) \to E_{k,h}^{(j),r}, r = 1,2,3, \quad (7)$$

Umumiy holda hujjatni fragmentlash va undan axborot birliklarini chiqarib olish jarayoning matematik modeli quyidagicha bo‘ladi:

$$H_h \xrightarrow{A_\varphi, \Psi_\varphi} X_i^{(j)} \xrightarrow{\Gamma^{(j)}, KSB} E_{k,h}^{(j),r}, r = 1,2,3 \qquad (8)$$

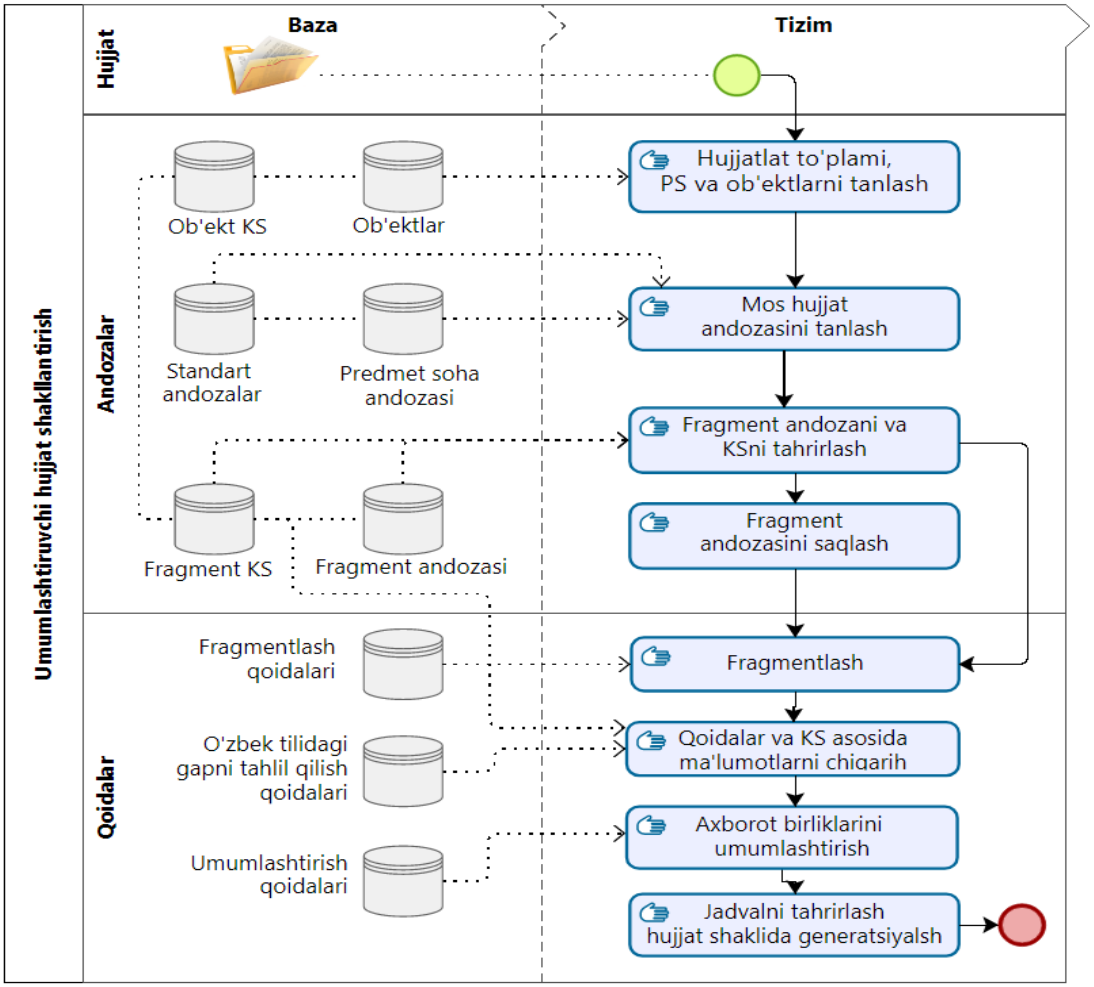
bu yerda $E^{(j),1}$ – axborot birligi nomi, $E^{(j),2}$ – axborot birligi qiymati, $E^{(j),3}$ – axborot formati, k – olingan axborot birliklari soni, (j) – fragment turi, h - joriy hujjatning tartib raqami.

Axborot massivlarini umumlashtirish. endi yuqorida aniqlagan hujjatlardan olingan axborot massivlarini umumlashtirish masalasi qaraladi. (8) formuladan $E_h^{(0)}$ - joriy tahlildagi h-hujjatning sarlavha fragmenti axborot nomi va axborot birliklari boshqa φ_j $(j = 1..4)$ fragment axborot birliklarini umumlashtirishda ishtirok qiladi. Shuni inobatga olib umumlashtirilgan hisobot hujjatni (Δ) shakllantirish modelini quyidagicha ifodalash mumkin:

$$\Delta = \cup_{h=1}^{nh} \cup_{j=1}^{4} \Lambda^{(j)} \left(E^0 \bowtie E_{k,h}^{(j),r} \right), \qquad (9)$$

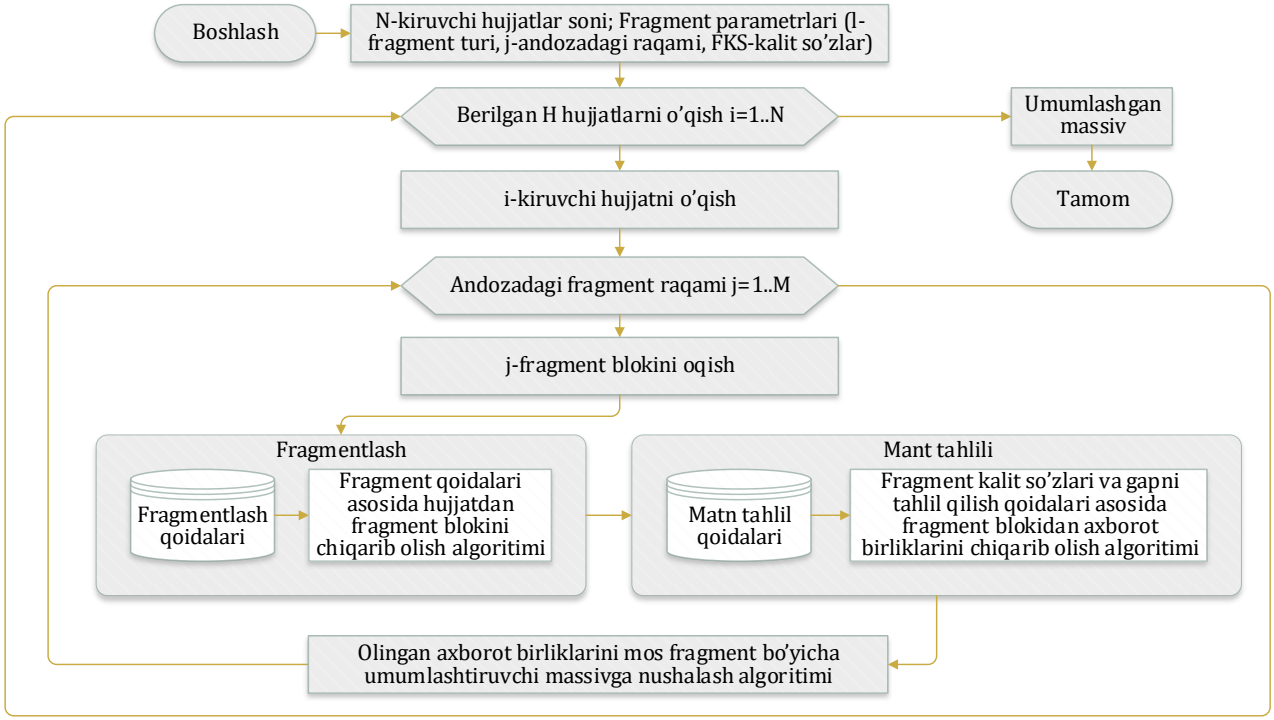
bu yerda (j)-fragment turi, nh-tahlildagi hujjatlar soni, $\Lambda^{(j)}$–mos (j)-fragment axborot birliklarini E^0 bilan birga umumlashtirish funksiyasi.

Λ umumlashtirish bosqichida muhim ahamiyat kasb etib, u har bir fragment uchun alohida **umumlashtirish qoidalariga** ega.



7-rasm**.** ARS-Uz tizimida axborot oqimi va jarayonlar bajarilishi

Axborot birliklarini chiqarish bosqichi ARS-Uz tizimining eng muhim algoritmik bo‘g‘ini bo‘lib, bu algoritm KS va qoidalar bazasi orqali matndan KSga mos axborot birliklarini chiqaradi. Natija maxsus axborot massivlariga mos fragment turi bo‘yicha yozib boriladi. Bunda umumlashtiruvchi massivning fragment turlariga mos axborot birliklarini yozish fragment umumlashtirish qoidalari, yani (9) model orqali amalga oshiriladi. Algoritmning qisqartirilgan blok sxemasi 8-rasmda berilgan.

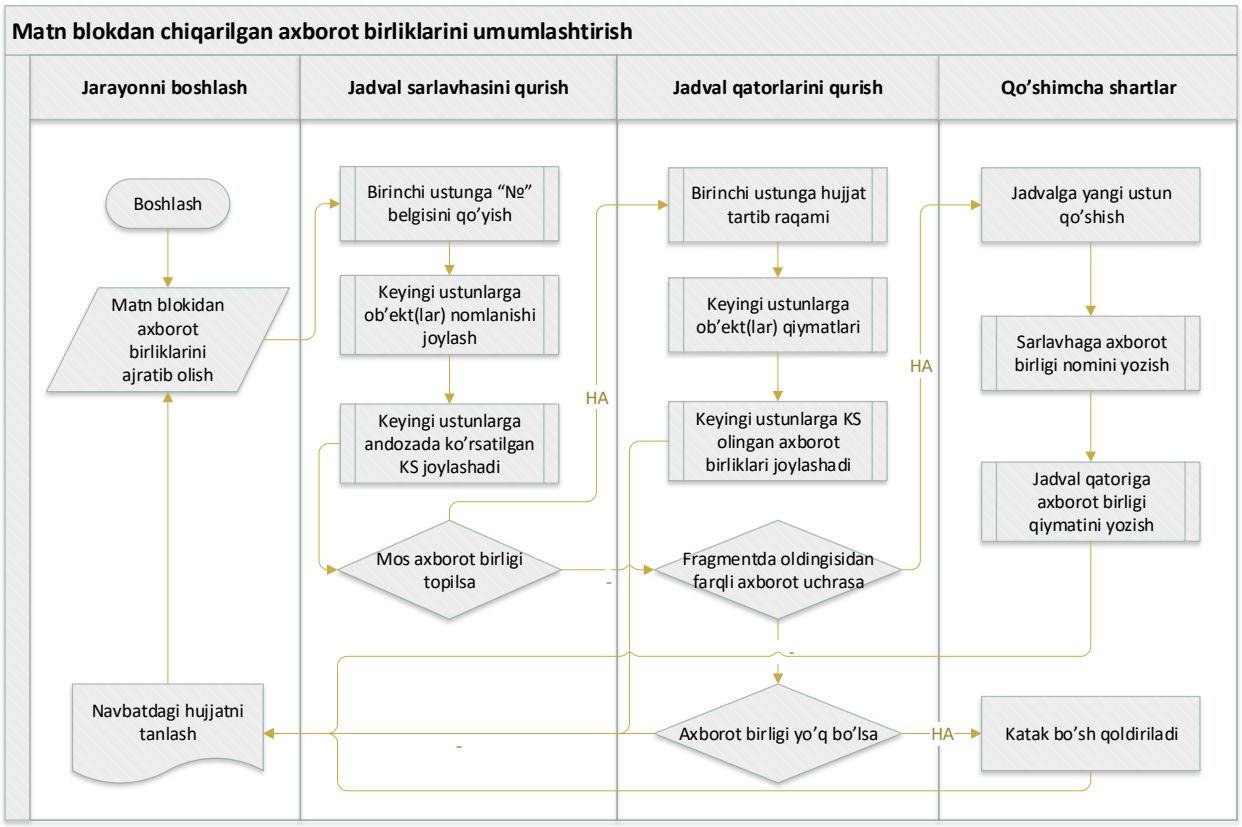


8- rasm**.** Hujjatlar fragmen blokidan axborot chiqarish algoritimi

Umumlashtirish bosqichida oldingi bosqichdan kelgan natijaviy jadvalli hujjat formallashtiriladi. Bunda asosan 3 ta bo‘lim ketma-ket va/yoki zarur holatda teskari qaytish amallari bajariladi: 1) avtomat ma’lumotlarni toifalash; 2) mexanik tahrirlash; 3) hujjatni generatsiyalash.

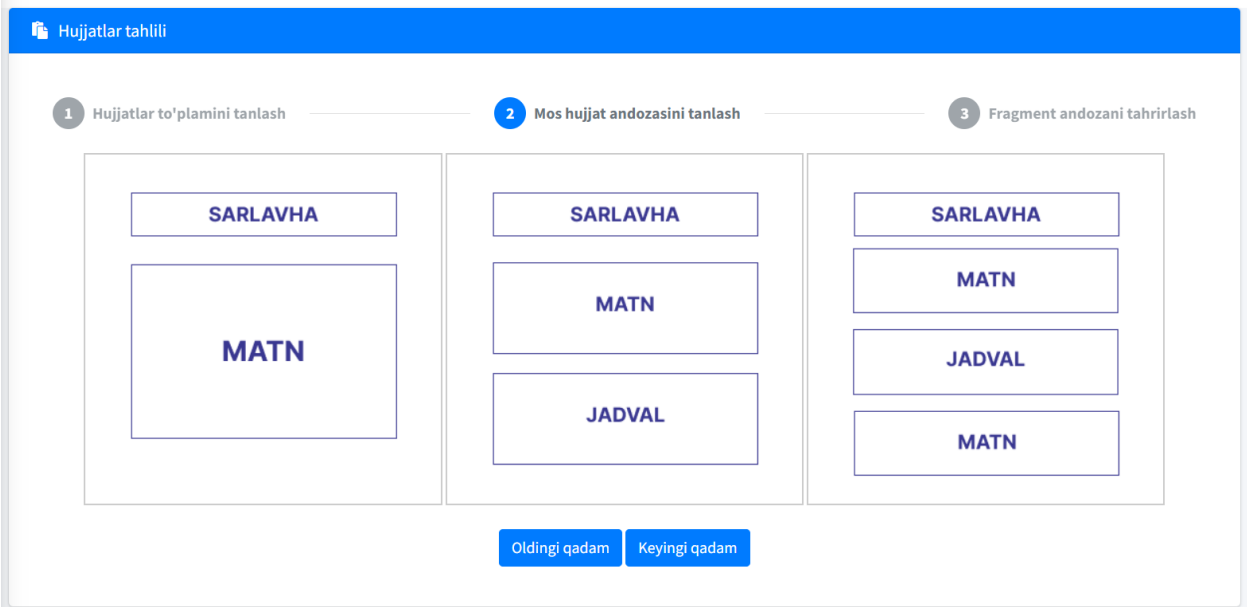
Ushbu bobda doimiy rivojlanishda bo‘lgan matnli hujjatlarni avtomatik qayta ishlash yo‘nalishlari hisoblangan KS ajratib olish, matnni umumlashtirish va nomlangan obyektni tanish uslubiyatlari tadqiq qilindi. O‘zbek tilidagi matnli hujjatlardan axborot birliklarini chiqarish va umumlashtirishning matematik modeli va tizim infratuzilmasi taklif etildi.

Dissertatsiya ishining to‘rtinchi bobi **“Kalit so‘zlar asosida bir jinsli hujjatlardan ma’lumotlarni ajratib olish va umumlashtirish dasturiy majmuasi”** deb nomlangan bo‘lib, unda tadqiqot davomida erishilgan natijalar bo‘yicha axborot dasturiy majmua ishlab chiqiladi. Jumladan, tadqiqot ishini bajarish davomida tahlil qilingan ma’lumotlar va olingan natijalar asosida bir xil tuzulmaga ega matnli hujjatlardan kalit so‘zlar orqali muhim axborot birliklarini ajratib olish va bu ma’lumotlardan ehtiyojga ko‘ra umumlashtiruvchi hisobotlar shakillantirib berish imkonini beradigan dasturiy majmuani ishlab chiqish jarayoni va bosqichlari haqida to‘xtalib o‘tiladi.



9- rasm. Matn blokidan chiqarilgan axborot birliklarini umumlashtirish algoritmining blok sxemasi.

Tahlil uchun ma'lumotlarni kiritish jarayoni 3 ta qadamda amalga oshiriladi. Birinchi qadamda foydalanuvchining mazkur andoza uchun nom tanlashi, fayllarni ko'rsatishi, predmet sohani tanlashi va obyektlarni tanlashi talab etiladi. Tegishli ma'lumotlar kiritilgandan so'ng ikkinchi qadamga o'tiladi (10-rasm).



10-rasm. Hujjat tahlilini yaratish uchun ma'lumotlarni kiritishning ikkinchi bosqichi.

Xususan dasturiy vosita arxitekturasini loyihalash, undagi foydalanuvchilarning turlari va ularning huquqlarini belgilash, dasturiy majmua uchun ma’lumotlar bazasini loyihalash va dasturiy vosita modullarining vazifalari shuningdek o‘zaro ta’sirlashishi haqidagi ma’lumotlar keltiriladi.

XULOSA

“Kalitli komponentlardan foydalangan holda bir jinsli hujjatlardan ma’lumotlarni ajratib olishning algoritmik va dasturiy majmuasi” mavzusidagi dissertatsiya ishi bo‘yicha olib borilgan tadqiqotlar natijasida quyidagi xulosalar taqdim etildi:

1. Elektron hujjat tuzilishi va fragment andozalari, kalit so‘zlar va bilimlar bazasi loyhalashtirildi. Natijada elektron hujjatlar uchun dizaynlar va shablonlar yaratish, kalit so‘zlar bilan ishlashni optimallashtirish va keng ko‘lamli ma’lumotlar bazasini tuzish imkoniyati paydo bo‘ldi. Bu esa, foydalanuvchilarga moslashuvchan va kengaytirilishi mumkin bo‘lgan tizimni taqdim etishga olib keldi.

2. Hujjatlardan toifali fragment bloklarini aniqlash va bu blokdan kalit so‘zlar asosida qiymatli axborotlani ajratib olish jarayoni matematik modeli ishlab chiqildi. Buning natijasida, axborotni avtomatik ravishda qayta ishlash orqali vaqt tejalishi va inson xatosini kamaytirish, bu esa ayniqsa katta miqdordagi matn ma’lumotlarini qayta ishlashda bir qator imkoniyatlarni taqdim etdi.

3. O‘zbek tilida ifodalangan matnli hujjatlarni fragmentlash, fragment matnidan kalit so‘zlar asosida qiymatli axborotlarni ajratib olish va umumlashtirishning qoidalar bazasi ishlab chiqildi. Buning natijasida, o‘zbek tilidagi matnli hujjatlarni qismlarga bo‘lish va umumlashtirish qoidalari bazasi, katta hajmdagi matnlarni qayta ishlash va tushunishni sezilarli darajada osonlashtirib, yanada samarali muloqot qilish, bilimlarni tarqatish va umumlashtirilgan ma’lumotlar asosida qarorlar qabul qilish imkoni yaratildi;

4. Bir jinsli elektron hujjatlardan kalit so‘zlar asosida qiymatli axborotlarni ajratib olish va umumlashtirish algoritmlari ishlab chiqildi. Asosan bu algoritmlarning samarali ishlashi tanlangan kalit so‘zlarning aniqligi, hujjatlarning o‘xshashligi va algoritmning o‘zi qanchalik murakkab ekanligiga bog‘liq ekanligi muhimdir. Natijada, bu algoritmlarni turli sohalardagi hujjatlarni tahlil qilish, ular ustida tadqiqotlar olib borish va chuqur tahlillar yaratish imkoniyatini berdi;

5. Ishlab chiqilgan algoritmlar asosida amaliy masalalarni yechishga ko‘mak beruvchi dasturiy majmuasi yaratilgan va real amaliy masalalarga tatbiq etilgan. Natijada, tashkilotga keladigan elektron hujjatlardan toifali fragment bloklarini aniqlash va bu blokdan kalit so‘zlar asosida qiymatli axborotlarni ajratib olish jarayoni avtomatlashtirilgan holda qayta ishlash imkonini berdi. Bu esa soha xodimlarini ish yuklama hajmini 15% ga qisqartirdi hamda keladigan hujjatlarga qayta ishlov berish asnosida javob qaytarish tezkorligini oshirish orqali ish samaradorligini 15% oshirishga imkon berdi.

НАУЧНЫЙ СОВЕТ PhD.13/05.05.2023.T.162.01. ПО ПРИСУЖДЕНИЮ УЧЕНЫХ СТЕПЕНЕЙ ПРИ НУКУССКОМ ФИЛИАЛЕ ТАШКЕНТСКОГО УНИВЕРСИТЕТА ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ

НУКУССКИЙ ФИЛИАЛ ТАШКЕНТСКОГО УНИВЕРСИТЕТА ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ ИМЕНИ МУХАММАДА АЛ-ХОРАЗМИЙ

КЕНЖАЕВ ХАМДАМ БАЗАРБАЕВИЧ

АЛГОРИТМИЧЕСКИЙ И ПРОГРАММНЫЙ КОМПЛЕКС ДЛЯ ИЗВЛЕЧЕНИЯ ИНФОРМАЦИИ ИЗ ИДЕНТИЧНЫХ ДОКУМЕНТОВ С ИСПОЛЬЗОВАНИЕМ КЛЮЧЕВЫХ КОМПОНЕНТОВ

05.01.04 - Математическое и программное обеспечение вычислительных машин, комплексов и компьютерных сетей

АВТОРЕФЕРАТ ДИССЕРТАЦИИ ДОКТОРА ФИЛОСОФИИ (PhD) ПО ТЕХНИЧЕСКИМ НАУКАМ

Нукус-2024

Тема диссертации доктора философии (PhD) по техническим наукам зарегистрирована Высшей аттестационной комиссией при Кабинете Министров Республики Узбекистан за номером V2023.2.PhD/T.3636

Диссертация выполнена в Нукусском филиале Ташкентского университета информационных технологий.

Автореферат диссертации размещён на трёх языках (узбекском, русском, английском (резюме)) на веб-странице Научного совета (www.tatunf.uz) и на информационно-образовательном портале "ZiyoNet" (www.ziyonet.uz).

Научный руководитель:	**Нишанов Ахрам Хасанович** доктор технических наук, профессор
Оффициальные оппоненты:	**Джуманов Жамолжон Худайкулович** доктор технических наук, профессор
	Рахимбоев Хикмат Жуманазарович доктор философии технический наук (PhD), доцент
Ведущая организация:	**Бухарский государственный университет**

Защита диссертации состоится «15» июля 2024 г. в 14[00] часов на заседании Научного совета PhD.13/05.05.2023.T.162.01 по присуждению ученых степеней при Нукусском филиале Ташкентского университета информационных технологий. (Адрес: 230100, г. Нукус, ул. А.Досназарова, 74. Тел.: (99861) 222-49-10, e-mail: tatunf@ tatunf.uz).

С диссертацией можно ознакомиться в Информационно-ресурсном центре Нукусского филиала Ташкентского университета информационных технологий (регистрационный № 1). (Адрес: 230100, г. Нукус, ул. А.Досназарова, 74. Тел.: (99861) 222-49-10).

Автореферат диссертации разослан «15» июля 2024 года.
(реестр протокола рассылки № 1 «12» июля 2024 года.

Б.Т.Каипбергенов
Председатель Научного совета по присуждению ученых степеней, доктор технических наук, профессор

Р.И.Отениязов
Секретарь Научного совета по присуждению ученых степеней, доктор технических наук, профессор

К.К.Сейтназаров
Председатель научного семинара при Научном совете по присуждению ученых степеней, доктор технических наук, профессор

ВВЕДЕНИЕ (аннотация диссертации доктора философии (PhD)

Актуальность и востребованность темы диссертации Во всем мире, в результате развития ИКТ и сети Интернет в мире бумажные источники информации, удовлетворяющие потребности людей, в частности такие издания, как документы, газеты и журналы, преобразовываются в электронную форму, в результате которого особое значение стал придаваться вопросам управления большими объемами и потоками информации, их обработки и предоставления пользователю. В настоящее время ведущими компаниями и учеными развитых стран, в том числе таких, как США, Англия, Южная Корея, Германия, Япония, Россия, передовыми в IT-сфере намечено внедрение в практику современных решений обработки текстов на естественном языке с использованием искусственного интеллекта, машинного обучения и средств анализа больших данных. В связи с этим особое внимание уделяется дальнейшему совершенствованию специальной системы, программных средств и алгоритмов, предназначенных для естественной языковой обработки текстов из документов однородного набора, по ключевым словам, извлечения точных данных и суммаризации документов.

Во всеми известными учеными проводятся значимые исследования связанные с процессом выявления и выделения отдельных элементов из документов, имеющих сходный формат или структуру. Приоритетными в этом направлении являются исследования, в том числе по суммаризации ценной информации из базы документов одинакового формата и структуры, и разработке алгоритмо-математических ресурсов, формирующих смысл содержания большого документа. В связи с этим актуальными считаются задачи преобразования их в единый формат и удаления ненужной информации или исправления ошибок.

В республике ведутся достаточные научные исследования и реформы связанные с поиском специальных фраз или форматов, указывающих на начало соответствующего раздела в документе, обработкой текстов на естественном языке, пониманием контекста в сложных документах, определения смысла и выявлением отношений между разными частями текста, алгоритмами, с использованием машинного распознавания для повышения их точности с течением времени особенно в сложных или разнообразных наборах данных, изучением и выявлением соответствующей информации из каждого процесса экстракции, а после извлечения данных преобразованием их в соответствующий формат, подходящий для анализа или интеграции в другие системы, такие как базы данных или средства анализа данных, а также с исследования посвященных интеграции с существующими базами данных, системами управления контентом или бизнес-процессами в целях реальной эффективности систем экстракции. В частности, в сфере цифрового формата управления и систематической обработки потока документов в органах государственного управления Республики Узбекистан обозначены решения таких задач, как “...Электронный обмен документами...

цифровизация операционных процессов...”[1] , “...информационная система для автоматизации образовательного процесса ... электронные формы...”, “... автоматизация процесса управления деятельности Министерства народного образования и его региональных подразделений...”[2]. При реализации этих задач важными являются проведение научно-практических исследований и разработка программных комплексов по обработке текстовых документов на узбекском языке, в частности по машинной суммаризации однотипных шаблонных документов и подготовке автоматических отчетов.

Настоящее диссертационное исследование в определенной степени служит реализации задач, указанных в законах Республики Узбекистан “Об электронном правительстве” (2015г.) и “Об обращениях физических и юридических лиц” (2017г.), Указах Президента Республики Узбекистан №УП-6079 “Об утверждении стратегии “Цифровой Узбекистан-2030” и мерах по ее эффективной реализации” от 5 октября 2020 года, №УП-4947 “О Стратегии действий по дальнейшему развитию Республики Узбекистан” от 7 февраля 2017 года, №УП-6097 “Об утверждении Концепции развития науки до 2030 года” от 29 октября 2020 года, Постановлении Президента Республики Узбекистан №ПП-4996 “О мерах по созданию условий для ускоренного внедрения технологий искусственного интеллекта” от 17 февраля 2021 года, Постановлениях Кабинета Министров Республики Узбекистан №7 “Об утверждении типового положения о порядке работы с обращениями физических и юридических лиц в органах самоуправления граждан” от 5 января 2018 года и №341 “Об утверждении Типового положения о порядке работы с обращениями физических и юридических лиц в государственных органах, государственных учреждениях и организациях с государственным участием” от 7 мая 2018 года, и в других нормативно-правовых актах, связанных с данной сферой.

Соответствие исследования приоритетным направлениям развития науки и технологий республики. Данное исследование выполнена в соответствии с IV приоритетным направлением развития науки и технологий республики “Информатизация и развитие информационно-коммуникационных технологий”.

Степень изученности проблемы. Ряд зарубежных ученых, в том числе А. О. Шигаров, И. В. Бычков, С.О.Шереметьева, А.Д.Усталов, Е.В. Стожок, А.Е.Хмельнов, Е.Ю.Хрустапев, А.В.Соловьев, Т.Г.Пенкова, H.H. Chen, S.C.Tsai, R.Campos, V.Mangaravite, A.Pasquali, A.Jorge, C.Nunes, A.Jatowt, Z.Jingsheng[3] и др., внесли свой вклад в решение таких задач, как

[1]Указ Президента Республики Узбекистан от 05.10.2020 г. №УП-6079 Об утверждении Стратегии “Цифровой Узбекистан-2030” и мерах по ее эффективной реализации.

[2] Указ Президента Республики Узбекистан от 29.04.2019 г. №УП-5712 Об утверждении концепции развития системы народного образования Республики Узбекистан до 2030 года.

[3] Christopher Manning, который связан со Стэнфордским университетом и много работал в области обработки естественного языка и его книга “Основы статистической естественной обработки языка” считается очень популярной; Jurafsky & Martin, их книга “Обработка речи и языка” распространено по всему миру; Andrew Ng, чья основная работа была посвящена глубокому распознованию и его применению в различных областях, но его вклад в область машинного распознавания оказал косвенное влияние на методы извлечения информации; Jacob Devlin, Sebastian Riedel, Yoshua

алгоритмический и программный комплекс извлечения информации из однородных документов с использованием ключевых компонентов.

Разработке и совершенствованию методов извлечения данных из однородных документов с использованием ключевых компонентов посвящены научные работы известных ученых Узбекистана. Из них стоит отметить труды М.М.Камилова, Ш.Х.Фазилова, Р.Н.Хамдамова, Н.С.Маматова, С.С.Раджабова, М.Х.Худойбердыева, А.Хамроева, А.Х.Нишанова и др.

В результате проведенных научных исследований достигнуты значительные результаты в решении задач практического применения алгоритмического и программного комплекса извлечения данных из однородных документов с использованием ключевых компонентов. В то же время теоретические и практические проблемы направления алгоритмического и программного комплекса извлечения информации из однородных документов с использованием ключевых компонентов недостаточно изучены.

Взаимосвязь диссертационного исследования с планами научно-исследовательских работ высшего учебного заведения. Диссертационная работа выполнена в рамках научно исследовательских проектов №22/19-Ф “Разработка и внедрение системы iGov-консультирование-обсуждение-мониторинг для повышения эффективности использования услуг национальной среды электронного правительства” (2019-2020), №333-U “Создание электронной галереи “Онлайн-музей” в системе художественного образования Узбекистана” (2021-2022 гг.), №IL-392103072 “Создание мобильного приложения для электронного управления животноводческими комплексами” (2022-2023 гг.) в соответствии с научно научно-исследовательскими планами Ташкентского университета информационных технологий имени Мухаммада аль-Хоразмий, Национального института художеств и дизайна имени Камолиддина Бехзода и Нукусского филиала Ташкентского университета информационных технологий имени Мухаммада аль-Хоразмий соответственно.

Цель исследования. Целью исследования является разработка алгоритмического и программного комплекса выделения ключевых информационных единиц из текстовых документов с одинаковой структурой по ключевым словам и формирования из этих данных суммарных отчетов по мере необходимости.

Задачи исследования: Задачами исследования являются разработка математической модели задачи суммаризации информационных единиц документа на основе ключевых слов и инфраструктуры программного комплекса;

проектирование шаблонов электронных документов, ключевых слов и базы знаний;

Bengio, Geoffrey Hinton, and Yann LeCun, которые в настоящее время считаются известными учеными в области глубокого распознавания, широко используемого при извлечении информации из текстов.

разработка методики извлечения и суммаризации ценной информации на основе шаблонов, фрагментов, ключевых слов электронных документов, и программного обеспечения;

разработка программного комплекса, способствующего решению практических задач в разрезе разработанных алгоритмов.

В качестве **объекта исследования** были взяты подходы, основанные на поиске, анализе и суммаризации необходимой информации, удовлетворяющей потребности пользователя, из большого объема информационных ресурсов.

Предмет исследования состоит из алгоритмов суммаризации большого количества документов, поступающих из систем электронного документооборота в результате исходящего запроса документа, или алгоритмов извлечения необходимых единиц информации из набора смешанных текстовых документов по заданному шаблону и ключевым словам, и программного комплекса, помогающий принимать решения.

Методы исследования. В ходе исследования были использованы методы интеллектуального анализа данных, обработки больших объемов информации и теории распознавания образов.

Научная новизна исследования состоит в следующем:

на основе шаблонов фрагментов, ключевых слов и базы знаний электронного документа спроектированы методы обработки текста;

разработана математическая модель процесса определения блоков категориальных фрагментов из документов и извлечения ценной информации из этих блоков на основе ключевых слов;

разработана база правил фрагментации текстовых документов, выраженных на узбекском языке, извлечения и суммаризации ценной информации на основе ключевых слов из текста фрагмента;

на основе базы правил извлечения и суммаризации ценной информации из текста фрагмента идентичных электронных документов разработаны алгоритмы извлечения и обобщения ценной информации.

Практические результаты исследования заключаются в следующем:

обоснована возможность улучшения процессов обобщения документов, принятия решений, составления аналитической отчетности в отраслях имеющих дело с объемной документацией;

разработанная автоматизированная система, имеющая возможность обработки гораздо большего объема документаций по сравнению с ручной обработкой, предназначена для повышения деятельности организаций, регулярно и оперативно обрабатывающие данные;

еще более доступность соответствующей информации, за счет эффективного извлечения основных структурных частей из документов, обосновала возможность поддержки процессов управления знаниями и поиска информации;

такие системы могут быть адаптированы к определенным областям или типам документов, что делает их универсальными инструментами для различных приложений. Например, они могут быть адаптированы для

получения информации о пациентах в области здравоохранения или деталей транзакций в банке;

Подводя итог, можно сказать, что алгоритмический и программный комплекс для извлечения информации из однотипных документов имеет огромное практическое значение благодаря своей способности повышать эффективность, точность и скорость обработки информации в различных секторах, предлагая при этом возможности масштабирования и адаптации для удовлетворения различных потребностей.

Достоверность результатов исследования. Достоверность общетеоретических выводов, отраженных в конце исследования объясняются такими факторами как точность системы, автоматически извлекающей соответствующую информацию из набора документов аналогичной структуры, сбор и обработка данных, обобщение однотипной документальной информации, проектирование, разработка и тестирование алгоритмов, лучшие практики кодирования программного обеспечения, модульность и масштабируемость, проверка, тестирование, анализ ошибок и улучшение.

Научно-практическая значимость результатов исследования. Научная значимость результатов исследования объясняется извлечением соответствующей информации из текстов на естественном языке в наборе аналогичных документов путем выявления и анализа основных компонентов, совершенствованием методов машинного обучения, сравнением, эффективностью суммаризации идентичных документов и критериями точности.

Практическая значимость результатов исследования заключается в автоматизации процессов извлечения данных, повышении точности поиска информации, применении в различных областях, улучшении систем управления знаниями, облегчении анализа контента, включении крупномасштабного анализа данных, масштабируемости и эффективности, гибкости, а также во вкладе в область компьютерной лингвистики, разработки программной системы, обеспечивающей современную поддержку специалистов отрасли.

Внедрение результатов исследований. На базе программного комплекса “ARS-Uz”, созданного на основе алгоритмов, разработанных в ходе научных исследований:

информационная система, разработанная на основе правил фрагментации текстовых документов на узбекском языке, извлечения и суммаризации ценной информации на основе ключевых слов из текста фрагмента, внедрена в отдел дошкольного и школьного образования Элликкалинского района. (Справка министерства дошкольного и школьного образования от 2 ноября 2023 г., за номером№ 01-03/4034). В результате на 21% сократилось время, затрачиваемое на такие процессы, как анализ и оперативный ответ на различные формы информации (письма, приказы и т.д.), поступающей от вышестоящих и нижестоящих организаций, а также обращения граждан. Это позволило повысить эффективность работы на 23%.

программный комплекс, созданный на основе построения математической модели структуры электронного документа и шаблонов фрагментов, проектирования ключевых слов и базы знаний, выявления категориальных фрагментных блоков из документов, а также процесса извлечения ценной информации из этого блока на основе ключевых слов был внедрен в отдел дошкольного и школьного образования Берунийского района (Справка министерства дошкольного и школьного образования от 2 ноября 2023 г., за номером № 01-03/4034). Данное внедрение позволило автоматизировать процесс выявления категориальных фрагментарных блоков из электронных документов, поступающих в организацию, и извлечения ценной информации из этого блока на основе ключевых слов. Это позволило снизить нагрузку на отраслевой персонал на 17%, а также повысить эффективность работы на 19% за счет повышения скорости реагирования на основе обработки входящих документов

программный комплекс, созданный на основе алгоритма извлечения и обобщения ценной информации по ключевым словам из электронных документов, внедрен в отдел дошкольного и школьного образования Амударьинского района (Справка министерства дошкольного и школьного образования от 2 ноября 2023 г., за номером № 01-03/4034). В результате организация создала для своих сотрудников такие возможности, как получение и автоматический анализ документов на основе базы данных ключевых слов, относящихся к 7 различным категориям, а также извлечение по запросу ценной информации из каждого документа. Это позволило повысить эффективность работы на 24% за счет сокращения времени на 22%.

Апробация результатов исследования. Основные теоретические и практические результаты диссертации были обсуждены на 4 международных и 7 республиканских научно-технических и научно-практических конференциях.

Публикация результатов исследования. Основные результаты по теме исследования опубликованы в 23 научных работах, из них 8 в научных изданиях, рекомендованых ВАК Республики Узбекистан для публикации основных результаты докторских диссертаций, в том числе 14 в республиканских и 6 в зарубежных журналах. При этом 3 опубликованы в других зарубежных журналах и 3 программных продукта, созданных для ЭВМ, получили регистрационные удостоверения Агентства интеллектуальной собственности при Министерстве юстиции Республики Узбекистан.

Структура и объем диссертации. Диссертация состоит из введения, четырех глав, заключения, списка использованной литературы и приложений. Объем диссертации составляет 110 печатных страниц.

ОСНОВНОЕ СОДЕРЖАНИЕ ДИССЕРТАЦИИ

Во **введении** диссертации дана кратко излагается информация об актуальности темы исследования, степени изученности проблемы, цели, задачах, объекте, предмете исследования, методах исследования и научной

гипотезе, основных положениях и научных новизнах выносимых на защиту, о теоретической и практической значимости, внедрении результатов и апробации работы, опубликованных результатах, используемых моделях и алгоритмах, области применения и этапы реализации

Первая глава диссертации называется **“Алгоритмические подходы к извлечению информации из документов”,** и в ней представлен анализ алгоритмов и методов извлечения информации из документов, т.е. формирования ключевых слов, оценки сходства текстов. Также представлен анализ подходов к распознаванию имен объектов по текстовым данным и извлечению таблиц из документов, а также автоматизированных методов обобщения текстов.

Алгоритмы извлечения ключевых слов из документов и методы оценки сходства текстов приобретают важное значение при решении задач обработки естественного языка (NLP), формирования ключевых слов (KS) в исследованиях и обобщения текстов на их основе, классификации и кластеризации текстов. В нем представлены алгоритмы извлечения ключевых слов из документов, классификация методов измерения взаимного сходства текстов, анализ методов оценки сходства ключевых слов человека и алгоритма.

Кроме того, в следующем параграфе проведен сравнительный анализ подхода распознавания именованного объекта (Named Entity Recognation-NER) из текстовых данных, распознавания именованного объекта на основе правил (NER) и существующих методов, а также представлены преимущества NER основанного на правилах. NER занимается идентификацией именованных объектов, таких как имена людей, организаций, время и местоположение, из данного набора данных или корпуса. К именованным объектам относятся объекты предметной области (медицинские, пищевые), именованные объекты, определенные в корпусе и т.п. Например:

Текст:
Хуршид выиграл грант в размере 25 000 долларов на обучение в Оксфорде в 2023 году.
Выход:
Хуршид [Человек]выиграл грант в размере 25 000 долларов на обучение в Оксфорде[Организация] в 2023 году[Время].

Существует три основных подхода NER: подход, основанный на словаре, подход, основанный на обучении (learning based), и подход, основанный на правилах (rule-based).

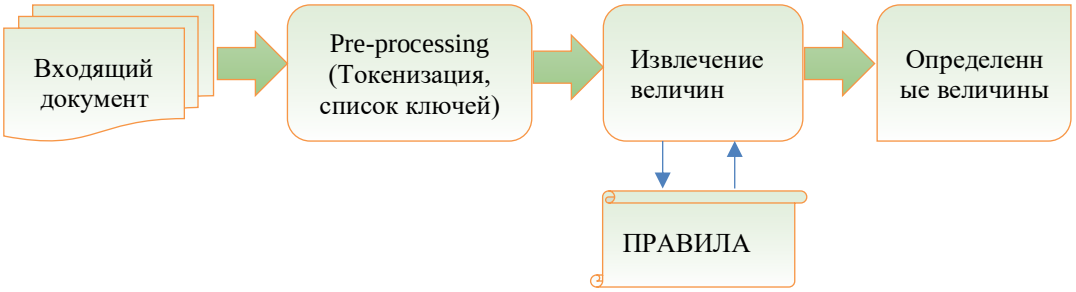


Рис. 1. Процесс количественного определения с использованием NER на основе правил

NER на основе словаря используется для извлечения информации из документов, поскольку он может предоставлять ID информацию по распознанным термам. Этот метод идентифицирует именованные объекты путем сопоставления термов. Подходы, основанные на словарном запасе, имеют такие ограничения, как ложноположительное распознавание и отсутствие единого ресурса для охвата недавно опубликованных названий. Хотя этот метод имеет высокий уровень точности, но он может распознавать только NER, включенный в словарь.

Хотя эти системы имеют некоторые ограничения, они не являются дорогими портативными и зависят от предметной области. Это требует знания предметной области и человеческого опыта для навыков программирования. NER системы, основанные на правилах, предназначены только для одной предметной области и не копируются в другие предметные области.

Преимущество NER, основанного на правилах, заключается в том, что эксперты могут создавать правила для каждой особенности естественного языка для конкретной области.

Хотя именованные объекты могут быть извлечены из текстов в документе с помощью NER, но оно не может быть применено для табличных данных в документе. По этой причине табличная модель извлечения информации из таблиц в документе будет рассмотрена в следующем разделе.

Предлагаемая табличная модель предназначена для представления фактов о таблицах в процессе логического заключения. Модель состоит из двух уровней:

Физический уровень описывает геометрические позиции, стили (графическое форматирование) и содержимое ячеек. Этот уровень $T_p = (S_r; S_c; C)$ состоит из следующих наборов: S_r – набор строк, а S_c – набор столбцов; $C - c = (c'; p; G)$ – набор ячеек, содержащий: c' - контент (значение); $p = (c_l; r_t; c_r; r_b)$ – координаты в S_r строках и S_c столбцах (c_l - левый столбец, r_t – верхняя строка, c_r – правый столбец и r_b – нижняя строка); G – набор настроек стиля (индикаторы шрифта, цвета, выравнивание текста, стили границ и т. д.);

Логический уровень представляет семантические отношения (т. е. пары ячейка-роль, заголовок-значение, заголовок-заголовок и заголовок-размер). Этот уровень $T_l = (D; L_r; L_c; E)$ состоит из следующих наборов: $D = \{D_i\}$ – набор размеров, представленных в переработанной таблице. Каждый из них представляет собой набор значений измерения $D_i = \{d_j\}$; L_r – дерево заголовков строк, а L_c – дерево заголовков столбцов. Эти деревья представляют отношения между их заголовками. Каждый заголовок имеет содержимое 1 $l = (l')$, l' не является значением измерений D_i: $: l' \notin \bigcup D_i$. E – $e = (e'; D'; L')$ набор записей: e' – содержание; D' – это набор значений измерения D_i , связанных с записью; L' – это набор заголовков из деревьев L_r и L_c, связанных с записью.

На этапе интерпретации извлечения таблиц из документов для анализа и интерпретации таблиц был предложен формальный язык правил, получивший

название CRL (Cells Rule Language). При этом рассматривался вопрос извлечения информации из произвольных полуструктурированных таблиц и загрузки их в базу данных с помощью стандартных инструментов ETL (Extract, Transform, Load). Предлагаемая схема процесса интерпретации неструктурированных табличных данных показана на рис. 2.

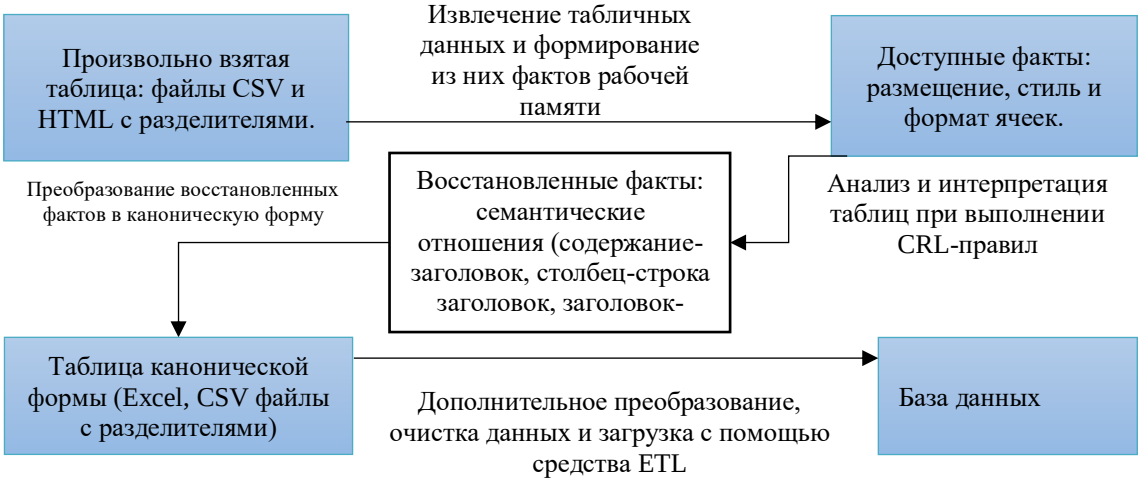


Рис. 2. Схема интерпретации неструктурированных табличных данных путем выполнения правил анализа и интерпретации таблиц.

Извлечение таблиц играет решающую роль в преобразовании неструктурированных данных в структурированный и практичный формат, делая данные более доступными, анализируемыми и важными для различных областей и различных приложений. В настоящее время реализовано несколько программных инструментов предназначенных для извлечения таблиц из документов.

Вторая глава диссертационной работы называется **“Методы и алгоритмы извлечения информации из идентичных документов”**, в которой исследуются направления обработки документов через информационные системы (автоматическое генерирование документа через программу и распознование документов автоматической системой) и способы представления в них информации. В частности, подробно рассмотрены шаблонные документы, содержимое документа и основы его фрагментации, генерация документов на основе системной базы данных или исходной таблицы, а также методы и алгоритмы машинного распознования электронных документов из внешних источников. В результате исследования предложены методы решения проблемы. Поэтому в данной работе подробно анализируются исследовательские работы, проведенные по трем вопросам (представление информации, автоматическое формирование и распознования документов системой), и предлагаются требования к системе, работающей с универсальными документами.

Основным вопросом в исследовании является извлечение соответствующих информационных единиц из набора текстовых документов и их обобщение в итоговый документ. Основная цель системы автоматической

суммаризации текста (ATS) – сохранить краткую информацию, содержащую основное содержание вводного документа, в меньшем пространстве и свести к минимуму дублирование.

Рис. 3. (а) однодокументный или (б) многодокументный автоматический суммаризатор текста.

В системах ATS существует множество классификаций. Об этих классификациях дан подробный систематический анализ. Системы ATS можно классифицировать в основном по следующим признакам (рис. 4).

Рис. 4. Классификация систем ATS

Основная классификации суммаризатора будет проводится по ее подходам (экстрактивным, абстрактным или гибридным). Подход суммаризации экстрактивного текста выбирает наиболее важные предложения во вводном документе (документах), и эти выбранные предложения объединяются в заключении. Подход суммаризации гибридного текста подход представляет собой сочетание экстрактивного и абстрактного подходов (рис.5).

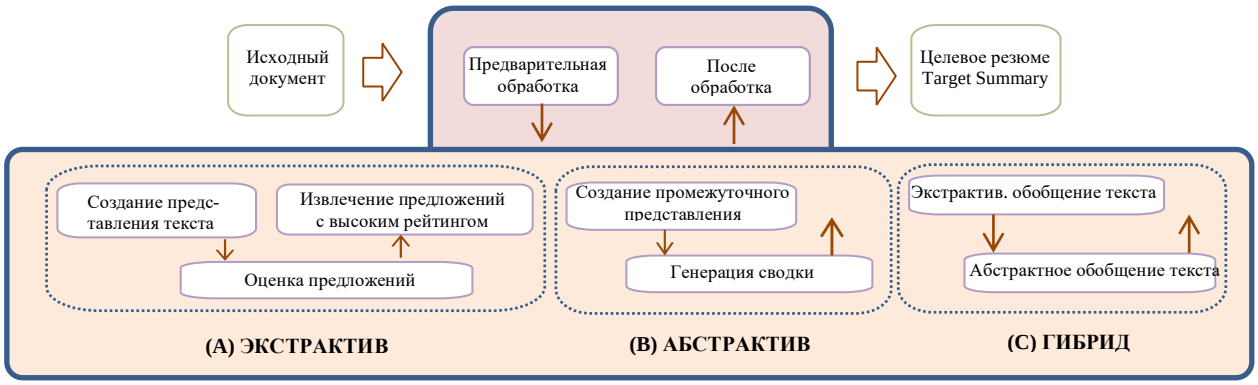


Рис. 5. Архитектура подходов к суммаризации текста

При проектировании и внедрении систем ATS используются различные компоненты и методы. В общем случае документ можно охарактеризовать как:

$$f(\alpha, \beta), \tag{1}$$

где α – матрица параметров документа, β – матрица описания структуры документа.

Документ представляет собой таблицу, размера $I \times k$, значений различных параметров α:

$$\alpha = \begin{bmatrix} p_{1,1} & \cdots & p_{1,I} \\ \vdots & \ddots & \vdots \\ p_{k,1} & \cdots & p_{k,I} \end{bmatrix}, \tag{2}$$

где p –параметр значения.

β представляет собой таблицу фрагментов документа размером $m \times n$, расположенных в определенном порядке и выражается следующим образом:

$$\beta = \begin{bmatrix} f_{1,1}(\Omega, \theta, \gamma) & \cdots & f_{1,m}(\Omega, \theta, \gamma) \\ \vdots & \ddots & \vdots \\ f_{n,1}(\Omega, \theta, \gamma) & \cdots & f_{n,m}(\Omega, \theta, \gamma) \end{bmatrix}, \tag{3}$$

где f – фрагмент документа, Ω –правило получения данных, θ – структура фрагмента (шаблон), γ – правило построения и визуализации (алгоритм). α и β взаимосвязаны через Ω, при этом некоторые элементы строки матрицы α являются элементами набора параметров Ω. Это позволяет изменять фрагменты документа, изменяя заданный массив данных. Таким образом, для набора параметров Ω малый набор элементов i-й строки матрицы равно M_i и правило получения массива данных имеет вид $\Omega(M_i)$, где $M_i \subset P_i$, $P_i \in \{p_{i,1} \dots p_{i,I}\}$.

Для фрагментов с одинаковой структурой, помимо массива данных, могут быть изменены параметры правил построения и визуализации. Когда функция параметров $\phi(y)$ вводится аналогично массиву данных, правило визуализации будет выглядеть следующим образом: $\gamma(\phi(y))$. Теперь матрица представления структуры документа задается формулой (4),

$$\begin{bmatrix} f_{1,1}(\Omega_j(M_i), \theta_g, \gamma_k(\phi(y))) & \cdots & f_{1,m}(\Omega_j(M_i), \theta_g, \gamma_k(\phi(y))) \\ \vdots & \ddots & \vdots \\ f_{n,1}(\Omega_j(M_i), \theta_g, \gamma_k(\phi(y))) & \cdots & f_{n,m}(\Omega_j(M_i), \theta_g, \gamma_k(\phi(y))) \end{bmatrix}, \tag{4}$$

при этом $i \in I$ – номер строки матрицы параметров документа, $g \in G$ – номер формы представления фрагмента, $j \in J$ – номер правила выбора заданных данных для построения фрагмента документа, $k \in K$ – номер правила визуализации фрагмента документа. Эта (4) формула обеспечивает соотвтествие ранее рассмотренных требований при разработке технологии, сохранении и реализации многих форм. В результате предложена схема работы программного средства формализации документов и разработана технология распознования и представления документов.

Также в этой главе разработана функциональная модель процесса генерации документов. Он состоит из четырех основных этапов: создание шаблона документа; создание запросов к базе данных; установка параметров заполнения шаблона и генерация документа в соответствии с шаблоном.

На этапе **создания шаблона** пользователем создается визуальный макет для каждого типа документа, на этапе создания **запросов к базе данных** подготавливаются данные для заполнения созданного шаблона, на этапе **настройки шаблона** элементы шаблона сравниваются с данными исходной таблицы или данными из запросов, а на этапе **генерации документа в соответствии шаблоном** введенные специальные поля заполняются автоматически согласно заданным пользователем параметрам и условиям.

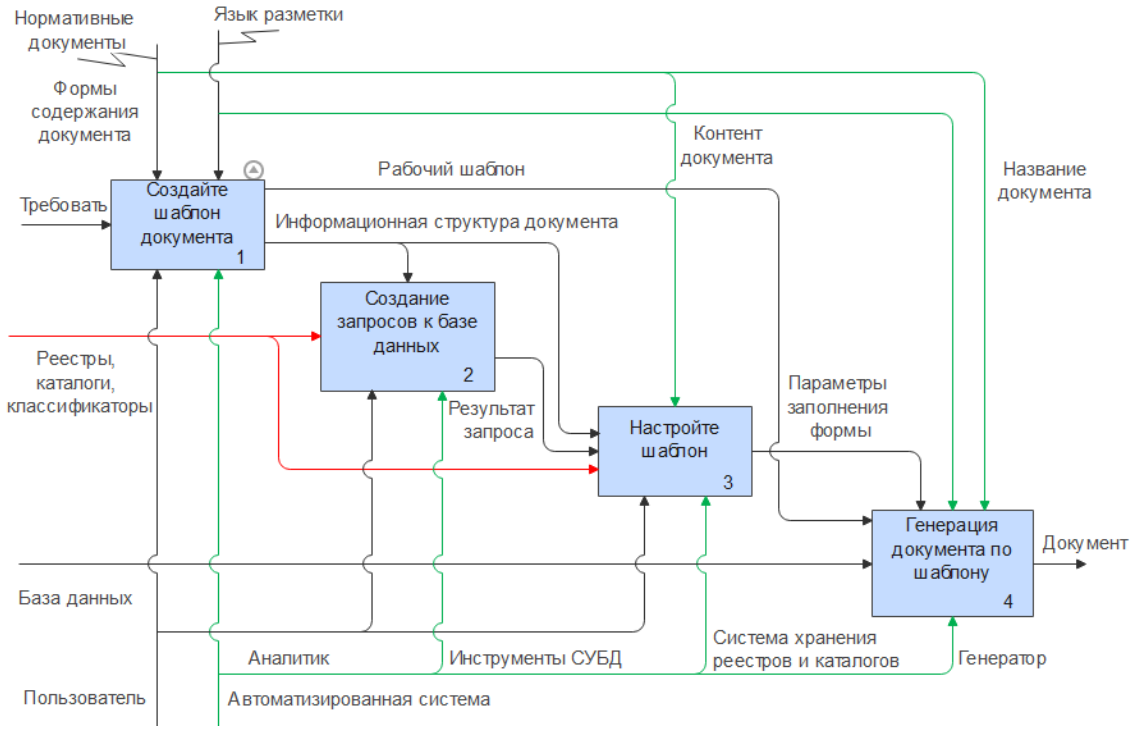


Рис. 6. Функциональная модель процесса формирования документов

Третья глава диссертационной работы называется **“Суммаризация документов, выраженных на узбекском языке, на основе ключевых компонентов”**, в которой изложена новая методика, математические основы суммаризации текстовых электронных документов, выраженных на узбекском языке, на основе компонентов ключевых слов, опираясь на исследованные

методики, в частности на механизмы и алгоритмы суммаризации ключевых слов, NER, ATS и таблиц.

Как видно из примера процесса суммаризации документа на основе человеческого интеллекта, для рассматриваемого вопроса необходимо построить **базу ключевых слов (KSB)**, представляющую виды объектов и их наименования, термы (наименования) относящиеся к предметной области, показатели. Этот KSB определяется как:

$$KSB = \{L, O, P, T, B, R, Q\}, \quad (5)$$

где:

$\boldsymbol{L} = \{l_1, l_2, \dots, l_{nl}\}$ – набор имен объектов (Кегейли, Нукус,…). Объект – существующая организация/территория с физическим/виртуальным адресом;

$\boldsymbol{O} = \{o_1, o_2, \dots, o_{no}\}$ – набор типов объектов (район, махалля, школа …). Имя объекта определяется из документа через тип объекта;

$\boldsymbol{P} = \{p_1, p_2, p_3, \dots., p_{np}\}$ – предметные области (сельское хозяйство, народное образование);

$\boldsymbol{T} = \{t_1, t_2, \dots, t_{nt}\}$ – набор термов на естественном языке (предприниматель, безработный, село, земля, книга, ученик, гражданин, посещаемость, пенсия, вода, лето,..);

$\boldsymbol{B} = \{b_1, b_2, \dots, b_{nb}\}$ – набор термов (комбинаций) и его синонимов, относящихся к типам объектов, которые можно назвать ключевыми словами объекта (посещаемость (присутствие, посещаемость урока…), ученик (студент, учащийся, слушатель,….), учитель (педагог, наставник, мастер, тренер, ….),...);

$\boldsymbol{R} = \{r_1, r_2, \dots, r_{nr}\}$ – набор цифр (0,…, 9, один, два, … десять, сто,тысячи, миллион, миллиард, I, V. X, C…).

$\boldsymbol{Q} = \{q_1, q_2, \dots, q_{nq}\}$ – набор сокращений (ТУИТ КГУ, УЗМУ, МДО, …)

Эти обозначения известны как KSB. Сначала через KSB определяются объект и его название, затем извлекаются единицы информации на основе термов.

Извлечение фрагментов из документа осуществляется через функцию:

$$\Theta\left(H_h, A_\varphi, \Psi_\varphi\right) \to X_i^{(j)}, \quad (6)$$

где H_h обозначает h-документ входящего набора, A – выбранный шаблон, Ψ – правило фрагментации, $X_i^{(j)}$ – блок выделенного (j)- фрагмента, i-появление этого типа фрагмента.

В общем случае функция извлечения информационных единиц выражается как:

$$F\left(X_i^{(j)}, \Gamma^{(j)}, FKS\right) \to E_{k,h}^{(j),r}, r = 1,2,3, \quad (7)$$

В общем случае математическая модель процесса фрагментации документа и извлечения из него информационных единиц будет выглядеть следующим образом:

$$H_h \xrightarrow{A_\varphi, \Psi_\varphi} X_i^{(j)} \xrightarrow{\Gamma^{(j)}, KSB} E_{k,h}^{(j),r}, r = 1,2,3 \quad (8)$$

где $E^{(j),1}$ – название единицы информации, $E^{(j),2}$ – значение единицы информации, $E^{(j),3}$ – формат информации, k – количество извлеченных единиц инофрмации, (j) – тип фрагмента, h – порядковый номер текущего документа.

Суммаризация информационных массивов. Теперь рассматривается вопрос суммаризации информационных массивов, полученных из выше указанных документов. Из Формулы (8) $E_h^{(0)}$ – информационное название и информационные единицы заголовочного фрагмента h -документа текущего анализа будут участвовать в суммаризации информационных единиц другого фрагмента φ_j $(j = 1..4)$.

Учитывая это, модель формирования суммарного отчетного документа (Δ) можно представить как:

$$\Delta = \bigcup_{h=1}^{nh} \bigcup_{j=1}^{4} \Lambda^{(j)} \left(E^0 \bowtie E_{k,h}^{(j),r} \right), \qquad (9)$$

где (j) – тип фрагмента, nh – количество документов в анализе, $\Lambda^{(j)}$– функция суммирования соответствующих единиц информации (j)-фрагмента вместе с E^0.

Λ приобретает важное значение на этапе суммаризации и имеет отдельные правила суммаризации для каждого фрагмента.

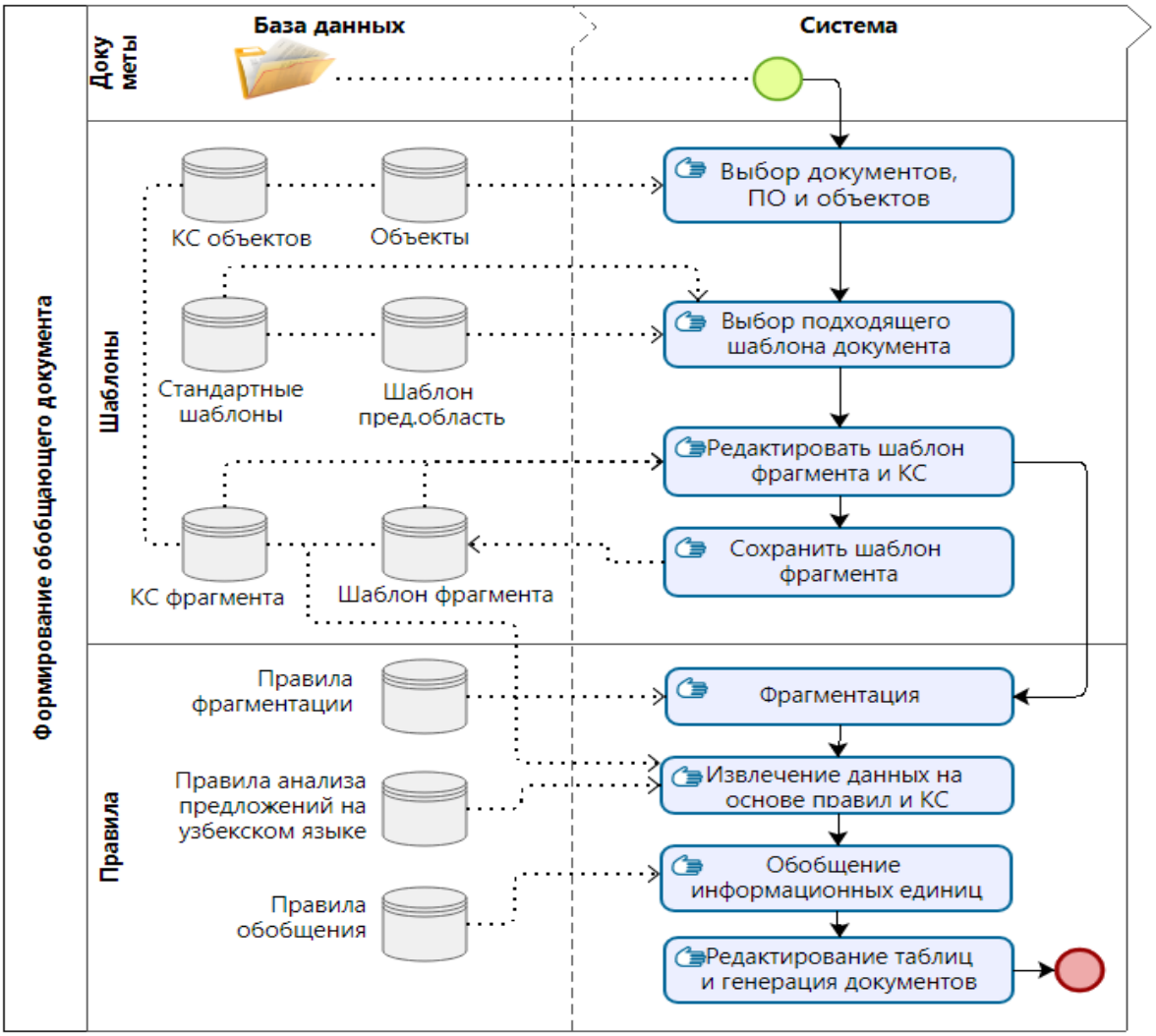


Рис. 7. Информационный поток и выполнение процессов в системе ARS-Uz

Этап вывода информационных единиц является важнейшим алгоритмическим звеном системы ARS-Uz, этот алгоритм выводит соответствующие к KS информационные единицы из текста через KS и базу правил. Результат записывается в специальные информационные массивы по типу соответствующего фрагмента. При этом запись единиц информации, соответствующих типам фрагментов суммарного массива, осуществляется с помощью правил суммаризации фрагментов, то есть с помощью модели (9). Сокращенная блок-схема алгоритма представлена на рис. 8.

Рис. 8. Алгоритм извлечения информации из блока фрагментов документов

На этапе суммаризации формализируется итоговый табличный документ с предыдущего этапа. При этом в основном выполняются следующие 3 последовательных и/или при необходимости обратных операции разделения: 1) автоматическая категоризация данных; 2) механическое редактирование; 3) генерация документа.

В данной главе исследованы методики извлечения KS, суммаризация текста и распознавания именованных объектов, которые считаются направлениями автоматической обработки текстовых документов, находящимися в постоянном развитии. Предложены математическая модель и инфраструктура системы для извлечения и суммаризации информационных единиц из текстовых документов на узбекском языке.

Четвертая глава диссертационной работы называется **“Программный комплекс извлечения и обобщения информации из идентичных документов на основе ключевых слов”,** в которой на основе результатов, достигнутых в ходе исследования, разработан информационный программный комплекс. В частности, рассмотрен процесс и этапы разработки программного комплекса, позволяющего на основе проанализированных в ходе

исследовательской работы данных и полученных результатов извлекать важные информационные единицы по ключевым словам из текстовых документов с однотипной структурой и по мере необходимости формировать из этих данных обобщающие отчеты.

Рис. 9. Блок-схема алгоритма суммаризации единиц информации, извлеченной из блока текста

Процесс ввода данных для анализа осуществляется в 3 этапа. На первом этапе пользователю необходимо выбрать имя для этого шаблона, указать файлы, выбрать предметную область и выбрать объекты. После ввода соответствующих данных переходится ко второму шагу (рис.10).

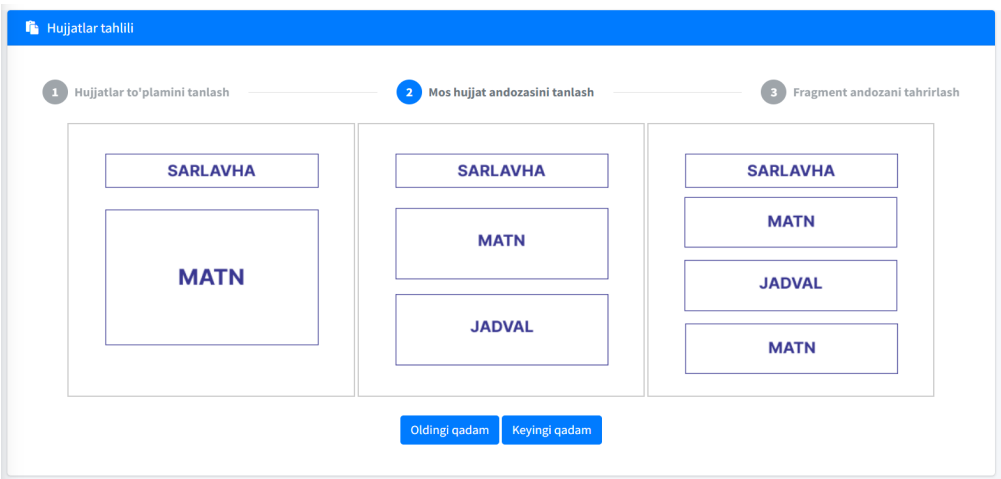


Рис. 10. Второй этап ввода данных для создания анализа документа

В частности, приводится информация о проектировании архитектуры программного средства, определении типов пользователей и их прав, проектировании базы данных программного комплекса, а также задачи и взаимодействие модулей программного инструмента.

ЗАКЛЮЧЕНИЕ

В результате исследования диссертационной работы на тему “Алгоритмический и программный комплекс для извлечения информации из идентичных документов с использованием ключевых компонентов” были представлены следующие выводы:

1. Спроектированы структура электронного документа и шаблоны фрагментов, ключевые слова и база знаний. В результате стало возможным создавать дизайны и шаблоны электронных документов, оптимизировать работу с ключевыми словами и создавать масштабную базу данных. В результате пользователям была предоставлена гибкая и расширяемая система.

2. Разработана математическая модель процесса выявления блоков категориальных фрагментов из документов и извлечения ценной информации из этого блока на основе ключевых слов. В результате достигается экономия времени и уменьшение человеческих ошибок за счет автоматической обработки информации, что дает ряд возможностей, особенно при обработке больших объемов текстовых данных.

3. Разработана база правил фрагментации текстовых документов, выраженных на узбекском языке, извлечения и суммаризации ценной информации из текста фрагмента на основе ключевых слов. В результате база правил разделения и суммаризации текстовых документов на узбекском языке существенно облегчила обработку и понимание больших объемов текстов, и позволила более эффективно общаться, распространять знания и принимать решения на основе обобщенной информации;

4. Разработаны алгоритмы извлечения и суммаризации ценной информации из идентичных электронных документов на основе ключевых слов. Важно, что эффективность этих алгоритмов зависит от точности выбранных ключевых слов, схожести документов и сложности самого алгоритма. В результате это позволило алгоритмам анализировать документы в различных областях, проводить их исследования и проводить глубокий анализ;

5. На основе разработанных алгоритмов создан и применен к реальным практическим задачам программный комплекс, помогающий решать прикладные задачи. В результате представлена возможность выявления блоков категориальных фрагментов из поступающих в организации электронных документов и автоматизированной обработки процесса извлечения ценной информации из этого блока на основе ключевых слов. Это снизило рабочую нагрузку сотрудников на 15% и позволило повысить эффективность работы на 15% за счет увеличения скорости реагирования при обработке входящих документов.

SCIENTIFIC COUNCIL AWARDING SCIENTIFIC DEGREES PhD.13/05.05.2023.T.162.01. AT NUKUS BRANCH OF THE TASHKENT UNIVERSITY OF INFORMATION TECHNOLOGIES

NUKUS BRANCH OF TASHKENT UNIVERSITY OF INFORMATION TECHNOLOGIES NAMED AFTER MUHAMMAD AL-KHWARIZMI

KENJAEV KHAMDAM BAZARBAEVICH

ALGORITHMIC AND SOFTWARE COMPLEX FOR EXTRACTING INFORMATION FROM IDENTICAL DOCUMENTS USING KEY COMPONENTS

05.01.04 - Mathematical and software support of computers, complexes and computer networks

ABSTRACT
DOCTOR OF PHILOSOPHY (PhD) DISSERTATIONS IN ENGINEERING SCIENCES

Nukus-2024

The topic of the doctor of philosophy (PhD) dissertation in technical sciences was registered by the Higher Attestation Commission under the Cabinet of Ministers of the Republic of Uzbekistan under the number V2023.2.PhD/T.3636

The dissertation has been prepared at Nukus branch of the Tashkent university of information technologies.

The abstract of the dissertation is posted in three languages (Uzbek, Russian, English (summary)) on the website of the Scientific Council (www.tatunf.uz) and on the information and educational portal "ZiyoNet" (www.ziyonet.uz).

Scientific supervisor:	**Nishanov Ahram Khasanovich** Doctor of Technical Sciences, Professor
Official opponents:	**Djumanov Jamoljon Xudayqulovich** Doctor of Technical Sciences, Professor
	Raximboev Xikmat Jumanazarovich Doctor of Philosophy in Technical Sciences (PhD), Docent
Leading organization:	**Bukharo state university**

The defense of the dissertation will take place "26" July 2024 at 14⁰⁰ hours at a meeting of the Scientific Council PhD.13/05.05.2023.T.162.01 at Nukus branch of the Tashkent University of Information Technologies. (Address: 230100, Nukus, A. Dosnazarov St., 74. Tel.: (99861) 222-49-10, e-mail: tatunf@ tatunf.uz).

The dissertation can be reviewed at the Information Resource Center of the Nukus branch of the Tashkent University of Information Technologies (is registered No. 1). (Address: 230100, Nukus, A. Dosnazarov street., 74. Ph.: (99861) 222-49-10).

Abstract of dissertation sent out on "15" july 2024 y.
(mailing protocol register No. 1 "12" july 2024 y.)

B.T.Kaipbergenov
Chairman of the Scientific Council for Awarding Academic Degrees, Doctor of Technical Sciences, professor

R.I.Oteniyazov
Scientific secretary of scientific council awarding scientific degrees, Doctor of Technical Sciences, professor

K.K.Seitnazarov
Chairman of the academic seminar under the scientific council awarding scientific degrees, Doctor of Technical Sciences, professor

INTRODUCTION (abstract of dissertation (PhD)

The purpose of the study. This paper focuses on the development of an algorithmic and software complex that can extract key information from text documents with a consistent structure using specific keywords and generate summarizing reports based on this extracted information as required.

The **object of the research** is the exploration of methods based on searching, analyzing, and summarizing necessary information from a large volume of information sources, specifically text documents, to fulfill the user's requirements.

Implementation of research results.

Based on the "ARS-Uz" software complex created on the basis of algorithms developed in scientific research:

The information system that developed on the basis of the rule base for fragmenting text documents expressed in Uzbek, extracting and summarizing valuable information from the fragment text was introduced to the Preschool and School Education Department of Ellikkala district.(Reference No. 01-03/4034 of the Ministry of Preschool and School Education of the Republic of Karakalpakstan dated November 2, 2023). As a result, the time required for processes such as analysis of various types of information (letters, orders, etc.) and citizens' appeals coming from the upper and lower levels of the organization and quick response has been reduced by 21%. This made it possible to increase work efficiency by 23%;

The software complex created on the basis of the construction of a mathematical model of the electronic document structure and fragment templates, key words and knowledge base, and the identification of categorical fragment blocks from documents and the process of extracting valuable information from this block based on key words was introduced to the Pre-school and School education department of Beruniy district. (Reference No. 01-03/4034 of the Ministry of Preschool and School Education of the Republic of Karakalpakstan dated November 2, 2023). As a result, the process of identifying category fragment blocks from the electronic documents received by the organization and extracting valuable information from this block based on keywords made it possible to work in an automated manner. This reduced the workload of industry employees by 17% and increased the efficiency of work by 19% by increasing the speed of response based on the processing of incoming documents;

The software complex, created on the basis of the algorithm for extracting and summarizing valuable information from electronic documents based on keywords, was introduced to the Pre-school and School education department of Amudarya district. (Reference No. 01-03/4034 of the Ministry of Preschool and School Education of the Republic of Karakalpakstan dated November 2, 2023). As a result, it created opportunities for the employees of the organization to receive documents based on the database of keywords belonging to 7 different categories and perform the automatic analysis process, as well as to extract valuable information from each document based on demand. This made it possible to increase work efficiency by 24% by reducing time by 22%.

The scientific novelty of the research.

fragment templates, keywords, and knowledge bases in the structure of electronic documents were designed according to text processing methods;

a mathematical model of the process of identifying categorical fragment blocks from documents and extracting valuable information from these blocks based on keywords was developed;

the rule base for fragmenting text documents expressed in the Uzbek language, extracting and summarizing valuable information from the fragment text based on keywords;

Algorithms for extracting and summarizing valuable information from the fragment text of homogeneous electronic documents were developed based on the rule base.

The structure and scope of the dissertation. The dissertation consists of an introduction, four chapters, a conclusion, a list of references, and appendices. The dissertation comprises 110 pages.

E’LON QILINGAN ISHLAR RO‘YXATI
СПИСОК ОПУБЛИКОВАННЫХ РАБОТ
LIST OF PUBLISHED WORKS

I bo‘lim (Часть I; Part I)

1. Nishanov A.X.,Kenjayev X.B. Hujjatlardan jadvallarni chiqarib olish masalasi, usullari va dasturiy ta’minotlar tahlili // Digital Transformation And Artificial Intelligence, Vol. 1 No. (2023), -B.148-157. (OAK Rayosatining 2023 yil 4 iyuldagi 340/5-son qarori)

2. Babadjanov E.S., Saidrasulov Sh.N., Kenjayev X.B. Tabiiy o‘zbek tildagi matnlarni formallashtirish orqali predmet sohasini aniqlash algoritmi // «Muhammad al-Xorazmiy avlodlari» ilmiy-amaliy va axborot-tahliliy jurnali. 2(24)/2023. -B.54-63. (05.00.00; №10)

3. Nishanov A.X., Babadjanov E.S., Kenjayev X.B., Avtomatik matnlarni umumlashtirish usullari tahlili // Journal of Advances in Engineering Technology. Information systems and processes. Vol.2 (10), April-Jun, 2023. B.37-46 DOI 10.24412/2181-1431-2023-2-37-46. (05.00.00; №12; Index Copernicus; ICV: 58.38)

4. Kenjayev X.B., Hujjatlardan kalit so‘zlarni chiqarish algorotmlari va matn o‘xshashligini baholash usullari tahlili // «Muhammad al-Xorazmiy avlodlari» ilmiy-amaliy va axborot-tahliliy jurnali. 3(25)/2023. -B.39-44. (05.00.00; №10)

5. Nishanov A.X., Kenjayev X.B. Ma’lumotlarni taqdim etish shakllari, hujjatlarni avtomat tanish va generatsiya qilish uslubiyatlari // «Muhammad al-Xorazmiy avlodlari» ilmiy-amaliy va axborot-tahliliy jurnali. 3(25)/2023. -B.64-73. (05.00.00; №10)

6. Kenjayev X.B., Elektron hujjatlarda jadvallar tuzilishini tanib olish // International Journal of Education, Social Science & Humanities. Finland Academic Research Science Publishers ISSN: 2945-4492. SJIF=7.502. Vol-11| Issue-7, 2023 https://doi.org/10.5281/zenodo.819520. -B-480-491. (05.00.00; №12; Index Copernicus; ICV: 58.38)

7. Babadjanov E.S., To‘liyev X.I., Kenjayev X.B. Mathematical model of summarization of important information units in text documents // International Conference on Information Science and Communications Technologies 2023. -B.-594-601. (OAK Rayosatining 2023 yil 29 avgustdagi 01-06/1410/55-son qarori)

II bo‘lim (Часть II; Part II)

1. Saidrasulov Sh.N., Kenjayev X.B. Analysis of methods and algorithms for keyword extraction // Universal journal of technology and innovation. 2023. Vol-1 ISSUE-5, ISSN 2992-8842, -B.1-19. https://doi.org/10.5281/zenodo.8422443.

2. Kenjayev X.B. Elektron hukumatda muhokama portalini ishlab chiqishga qo‘yilgan texnik va dasturiy talablar Muhammad al-Xorazmiy nomidagi Toshkent axborot texnologiyalari universiteti Nukus filiali Respublika ilmiy-amaliy anjuman materiallar to’plami. 2019 yil 29-30 oktyabr. –B. 347-350.

3. Kenjayev X.B. Elekrton hujjat almashish tizimlarining normativ-huquqiy asoslari // Muhammad al-Xorazmiy nomidagi Toshkent axborot texnologiyalari

universiteti Nukus filiali Respublika ilmiy-amaliy anjuman materiallar to’plami. Nukus-2016. 25-26 fevral.-B 194-195.

4. Kenjayev X.B. Texnik yo‘nalishdagi oliy ta’lim Microsoft office 365 AZURE texnologiyasining ahamiyati // Iqtisodiyotning real tarmoqlarini innovasiyon rivojlanishda axborot-kommunikasiya texnologiyalarining ahamiyati Respublika ilmiy-amaliy anjumanining ma’ruzalar to’plami. Toshkent-2017. 6-7 aprel. -B 62-64.

5. Kenjayev X.B. Oliy ta’lim muassalarining axborotlashgan muhitini yaratish // Iqtisodiyotning real tarmoqlarini innovasiyon rivojlanishda axborot-kommunikasiya texnologiyalarining ahamiyati Respublika ilmiy-amaliy anjumanining ma’ruzalar to’plami. Toshkent-2017. 6-7 aprel. -B 64-66.

6. Allanazarov A.B., Kenjayev X.B. Elektron hukumat xizmatlarining rivojlanish tahlili // Axborot-kommunikasiyalarning rivojlanish istiqbollari mavzusidagi Respublika ilmiy-amaliy anjuman ma’ruzalar to’plami. Qarshi-2018. 20-21 aprel.- B. 39-42.

7. Babadjanov E.S., Kenjayev X.B. Elektron identifikastiya va avtorizastiyalash infratuzilmasi // Axborot-kommunikasiyalarning rivojlantirish shoraitida innovasiyalar Respublika ilmiy-amaliy anjuman ma’ruzalar to’plami. Qarshi-2019. 15-17 aprel. –B. 20-22.

8. Babadjanov E.S., Kenjayev X.B. Elektron hukumat ishtirokchilari va o’zaro bog’liqlik modeli // «Iqtisodiyotning tarmoqlarini innovasion rivojlantirishda axborot-kommunikasiya texnologiyalarining ahamiyati» mavzusidagi Respublika ilmiy texnik anjumanining ma’ruzalar to’plami. 1-qism. Toshkent-2019. –B. 122-124.

9. Geldibayev B.Y., Kenjayev X.B. Ta'limda bulut texnologiyalarni qo'llash //«Iqtisodiyotning tarmoqlarini innovasion rivojlantirishda axborot-kommunikasiya texnologiyalarining ahamiyati» mavzusidagi Respublika ilmiy texnik anjumanining ma’ruzalar to’plami. 1-qism. Toshkent-2021 4-5 mart. –B. 146-147

10. Kenjayev X.B. Elektron hukumat tizimining imkoniyatlari va mavjud muammolarning tahlili // Nukus branch of Tashkent university of information technologies named after Muhammal al-Khwarizmi, “Matematik modellashtirish va axborot texnologiyalarining dolzarb masalalari” xalqora ilmiy-amaliy anjuman, Nukus-2-3 may, 2023 yil, 284-286 bet.

11. Kenjayev X.B., Toliev X.I. Ner yondashuvi bilan o’zbek tilidagi matndan miqdorlarni aniqlash qoidalari // International Journal of Advanced Technology and Natural Sciences ISSN: 2181-144X. DOI: 10.24412/2181-144X-2023-2-23-32. -B.23-32.

12. Kenjayev X.B. Bir jinsli matnli hujjatlardan axborotlarni ajratib olish algoritmi Muhammad al-Xorazmiy nomidagi Toshkent axborot texnologiyalari universiteti Nukus filiali Fan va ta’limni rivojlantirishda raqamli texnologiyalarning roli Respublika ilmiy-texnik anjumanining ma’ruzalar to‘plami. 2023 yil 28-29 noyabr. –B. 308-310.

13. A.X.Nishanov., Kenjayev X.B. Matnlarni kalit so‘zlar asosida umumlashtiruvchi tizimni yaratish vazifalari // “Xalq xo‘jaligi sohasida ilg‘or

texnologiyalar tadbiqi muammolari” mavzusidagi hududiy ilmiy-texnik konferensiyasi ma’ruzalar to‘plami. Nukus-27 dekabr 2023 yil -B 121-123.

14. Kenjayev X.B. O‘zbek tilidagi bir jinsli matnli hujjatlar asosida yarim avtomat hisobot tayorlovchi axborot tizimi // O‘zbekiston Respublikasi Adliya vazirligi huzuridagi Intellektual mulk agentligi. Ma’lumotlar bazasining rasmiy ro‘yxatdan o‘tkazilganligi to‘g‘risidagi guvohnoma № DGU 27316. 07.09.2023.

15. A.X.Nishanov., E.S.Babadjanov., Kenjayev X.B. ARS-Uz - matnli hujjalarni kalit so‘zlar asosida umumlashtirish tizimining ma’lumotlar bazasi // O‘zbekiston Respublikasi Adliya vazirligi huzuridagi Intellektual mulk agentligi. Ma’lumotlar bazasining rasmiy ro‘yxatdan o‘tkazilganligi to‘g‘risidagi guvohnoma №BGU 1142. 30.10.2023.

16. A.X.Nishanov, E.S.Babadjanov., Kenjayev X.B. Bir tuzilmali matnlardan kalitli komponentalar orqali umumlashtiruvchi jadval shakllantiruvchi dasturiy ta’minoti // O‘zbekiston Respublikasi Adliya vazirligi huzuridagi Intellektual mulk agentligi. Ma’lumotlar bazasining rasmiy ro‘yxatdan o‘tkazilganligi to‘g‘risidagi guvohnoma № DGU 28819. 03.11.2023.

Avtoreferat “Muhammad al-Xorazmiy avlodlari” ilmiy jurnali tahririyatida tahriridan o‘tkazildi hamda o‘zbek, rus va ingliz tillaridagi matnlarini mosligi tekshirildi. (№291 “05” 06. 2024 yil)