O‘ZBEKISTON MILLIY UNIVERSITETI
HUZURIDAGI ILMIY DARAJALAR BERUVCHI
DSc.03/30.12.2019.FM.01.02 RAQAMLI ILMIY KENGASH

O‘ZBEKISTON MILLIY UNIVERSITETI

TULIYEV ULUG‘BEK YULDASHEVICH

O‘ZBEK TILIDAGI HUJJATLARNING TEMATIK KLASSIFIKATORLARINI YARATISH

05.01.11 – Raqamli texnologiyalar va sun’iy intellekt

FIZIKA-MATEMATIKA FANLARI
bo‘yicha falsafa doktori (PhD) dissertatsiyasi
AVTOREFERATI

Toshkent – 2023

O‘UK: 519.95

Fizika-matematika fanlari bo‘yicha falsafa doktori (PhD) dissertatsiyasi avtoreferati mundarijasi

Оглавление автореферата диссертации доктора философии (PhD) по физико-математическим наукам

Content of dissertation abstract of doctor of philosophy (PhD) on physical-mathematical sciences

**O‘ZBEKISTON MILLIY UNIVERSITETI
HUZURIDAGI ILMIY DARAJALAR BERUVCHI
DSc.03/30.12.2019.FM.01.02 RAQAMLI ILMIY KENGASH**

O‘ZBEKISTON MILLIY UNIVERSITETI

TULIYEV ULUG‘BEK YULDASHEVICH

O‘ZBEK TILIDAGI HUJJATLARNING TEMATIK KLASSIFIKATORLARINI YARATISH

05.01.11 – Raqamli texnologiyalar va sun’iy intellekt

**FIZIKA-MATEMATIKA FANLARI
bo‘yicha falsafa doktori (PhD) dissertatsiyasi
AVTOREFERATI**

Toshkent – 2023

Fizika-matematika fanlari bo‘yicha falsafa doktori (Doctor of Philosophy) dissertatsiyasi mavzusi O‘zbekiston Respublikasi Oliy ta’lim, fan va innovatsiyalar vazirligi huzuridagi Oliy attestasiya komissiyasida B2022.4.PhD/FM831 raqam bilan ro‘yxatga olingan.

Dissertatsiya Mirzo Ulug‘bek nomidagi O‘zbekiston Milliy universitetida bajarilgan.

Dissertatsiya avtoreferati uch tilda (o‘zbek, rus, ingliz (rezyume)) Ilmiy kengash veb- sahifasi (http://ik-fizmat.nuu.uz/) va “Ziyonet” ta’lim axborot tarmog‘ida (www.ziyonet.uz) joylashtirilgan.

Ilmiy rahbar:	**Ignatev Nikolay Aleksandrovich** fizika-matematika fanlari doktori, professor
Rasmiy opponentlar:	**Matyaqubov Alisher Samandarovich** fizika-matematika fanlari doktori, dotsent
	Norov Abdusaid Murodovich texnika fanlari bo‘yicha falsafa doktori (PhD)
Yetakchi tashkilot:	**Urganch davlat universiteti**

Dissertatsiya himoyasi O‘zbekiston Milliy universiteti huzuridagi DSc.03/30.12.2019.FM.01.02 raqamli Ilmiy kengashning “___”_____ 2023-yil soat____ dagi majlisida bo‘lib o‘tadi (Manzil: 100174, Toshkent sh., Olmazor tumani, Universitet ko‘chasi, 4-uy. Tel.: (+99871) 227-12-24, faks: (+99871) 246-53-21, 246-02-24, e-mail: nauka@nuu.uz).

Dissertatsiya bilan O‘zbekiston Milliy universitetining Axborot-resurs markazida tanishish mumkin (_____ raqami bilan ro‘yxatga olingan). Manzil: 100174, Toshkent sh., Olmazor tumani, Universitet ko‘chasi, 4-uy. Tel.: (+99871) 246-02-24.

Dissertatsiya avtoreferati 2023-yil “____” _____kuni tarqatildi.
(2023-yil “____” _____ dagi _____ raqamli reyestr bayonnomasi).

M.M.Aripov
Ilmiy darajalar beruvchi ilmiy kengash raisi, f.-m.f.d., professor

Z.R.Raxmonov
Ilmiy darajalar beruvchi ilmiy kengash ilmiy kotibi, f.-m.f.d.

D.T.Muxamediyeva
Ilmiy darajalar beruvchi ilmiy kengash huzuridagi ilmiy seminar raisi, t.f.d., professor

KIRISH (falsafa doktori (PhD) dissertatsiyasi annotatsiyasi)

Dissertatsiya mavzusining dolzarbligi va zarurati. Jahon miqyosida olib borilayotgan ko‘plab ilmiy-amaliy tadqiqotlar, hozirgi kunda matnlarni tahlil qilishning (text mining) usullarini ishab chiqish va tabiiy tilga ishlov berish kabi masalalar bilan bog‘liq bo‘lmoqda. Bunday turdagi tadqiqotlar butun dunyo bo‘ylab keng miqyosda yoyilishga ulgurgan muammolardan biridir. Insonlar o‘rtasida bilimlarni yetkazishdagi eng kuchli vosita til hisoblanadi. Bu o‘rinda tabiiy tilning sintaktikasi va semantikasidagi farqlanishlar bilimlarni uzatishdagi jiddiy to‘siqlardan biridir. Sintaktika va semantikani shakllantirish hamda o‘zlashtirish uchun turli metodlardan foydalaniladi. Hujjatlarni tabiiy tilda saqlash uchun yirik hajmdagi berilganlar bazasini (korpuslar) yaratish yo‘li bilan uning mazmunini o‘rganishga harakat qilinadi. Hujjatlar tavsifidan bilimlarni ajratib olish jarayoni, berilganlarning intellektual tahlili (data mining) kabi jadal rivojlanayotgan tadqiqot sohalari doirasidagi qator muammolarning yechimidan iborat bo‘ladi. Shu sababli tabiiy tildagi matnli hujjatlarni sonli vektor ko‘rinishida ifodalash, ularning semantik bog‘lanishlari asosida tematik modellarini ishlab chiqish berilganlarning intellektual tahlilining muhim vazifalaridan biri bo‘lib qolmoqda.

Hozirgi kunda jahonda mashina lingvistikasi bo‘yicha juda katta hajmdagi ilmiy ishlar yig‘ilgan bo‘lib, amaliyotda ular tabiiy til semantikasi bilan ishlash metodlari va algoritmlarini qo‘llash imkonini beradi. Tabiiy tilning modellarini ishlab chiqish uchun uning xususiyatlarini hisobga olishga to‘g‘ri keladi. Tabiiy tilda taqdim etilgan Big Data muammosini yechish uchun usullar, algoritmlar va uskunaviy vositalar doirasida original natijalar qo‘lga kiritilgan. Aynan katta hajmdagi berilganlar yangi ilmiy bilimlar va semantik texnologiyalar rivoji bilan bog‘liq bo‘lgan texnologik yechimlar paydo bo‘lishida katolizatorga aylanib ulgurgan. Xususan, tabiiy tilga ishlov berish metodlari, uning turli ko‘rinishda aks etishidagi ko‘p ma’nolilikning hal etilishi maqsadli ilmiy tadqiqotlardan hisoblanadi.

Mamlakatimizda mustaqillik yillarida fundamental va amaliy ahamiyatga ega bo‘lgan ilmiy yo‘nalishlarga katta e’tibor qaratilmoqda. Sun’iy intellektni mashina yordamida o‘qitish uchun katta hajmda davlat tilidagi raqamli ma’lumotlarni shakllantirish, shuningdek, davlat tilidagi nutqni tahlil va sintez qilishni qo‘llovchi dasturiy mahsulotlarni ishlab chiqish ustuvor yo‘nalishlardan etib belgilandi[1]. Ushbu vazifalar predmet sohalar uchun lug‘atlar qurishni avtomatlashtirish, hujjatlarning mazmuniy haqqoniyligini hisoblash, semantik bog‘langanlik kabi masalalarning yechimlarini zaruratga aylantirdi. Matematik lingvistikaning dolazarb vazifasi, mashinali o‘rgatish metodlarini yaratishda foydalanish uchun predmet sohalarning o‘zbek tilidagi glossariy, tezaurus va lug‘atlar ko‘rinishidagi ontologiyasini ishlab chiqish hamda foydalanish hisoblanadi. Shuningdek, O‘zbekiston Respublikasini yanada rivojlantirish bo‘yicha Harakatlar strategiyasi asosida ilmiy tadqiqot va innovatsiya yutuqlarini amaliyotga joriy etish mexanizmlaridan iqtisodiyot tarmoqlarining samaradorligini oshirishda foydalanish muhim ahamiyatga ega hisoblanadi.

[1] O‘zbekiston Respublikasi Prezidentining 2021-yil 17-fevraldagi “Sun’iy intellekt texnologiyalarini jadal joriy etish uchun shart-sharoitlar yaratish to‘g‘risida”gi PQ-4996-sonli qarori.

O‘zbekiston Respublikasi Prezidentining 2017-yil 7-fevraldagi PF-4947-sonli “O‘zbekiston Respublikasini yanada rivojlantrish bo‘yicha harakatlar strategiyasi to‘g‘risida”gi Farmoni, 2017-yil 17-fevraldagi PQ-2789-sonli “Fanlar akademiyasi faoliyati, ilmiy tadqiqot ishlarini tashkil etish, boshqarish va moliyalashtirishni yanada takomillashtirish chora-tadbirlari to‘g‘risida”, 2017-yil 20-apreldagi PQ-2909-sonli “Oliy ta’lim tizimini yanada rivojlantirish chora-tadbirlari to‘g‘risida”gi qarori, 2018-yil 27-apreldagi PQ-3682-sonli “Innovatsion g‘oyalar, texnologiyalar va loyihalarni amaliyotga joriy qilish tizimini yanada takomillashtirish chora-tadbirlari to‘g‘risida”gi qarori, 2019-yil 24-mayda O‘zbekiston Milliy universitetidagi fan va ta’lim vakillari bilan uchrashuvdagi ma’ruzasi, 2020-yil 20-oktyabrdagi PF-6084-sonli “Mamlakatimizda o‘zbek tilini yanada rivojlantirish va til siyosatini takomillashtirish chora-tadbirlari to‘g‘risida”gi farmoni, 2021-yil 17-fevraldagi PQ-4996-sonli “Sun’iy intellekt texnologiyalarini jadal joriy etish uchun shart-sharoitlar yaratish to‘g‘risida”gi qarori hamda mazkur faoliyatga tegishli boshqa normativ-huquqiy hujjatlarda belgilangan vazifalarni amalga oshirishga ushbu dissertatsiya tadqiqoti muayyan darajada xizmat qiladi.

Tadqiqotning respublika fan va texnologiyalari rivojlanishining ustuvor yo‘nalishlariga mosligi. Dissertasiya respublika fan va texnologiyalar rivojlanishining IV. “Matematika, mexanika va informatika” ustuvor yo‘nalishi doirasida bajarilgan.

Muammoning o‘rganilganlik darajasi. Matematik tilshunoslik muammolarini hal qilishga talab bir tildan ikkinchi tilga tarjima qiluvchi tarjimon sifatida amaliyotda keng qo‘llaniladigan dasturiy ta’minotning mavjudligi bilan ham seziladi. Masalan, Google Translate tizimidan millionlab odamlar foydalanadi. Afsuski, tarjima sifatining natijalari hali ham professionallar qilishi mumkin bo‘lgan natijadan ancha uzoqdir. Tarjima semantikasini takomillashtirish, tabiiy tillardagi hujjatlarning tematik korpuslarini yaratish kabi masalalar bo‘yicha hali ko‘p ishlar qilinishi kerak. Yana bir muammo – tematik hujjatlar bilan ishlash jarayonini avtomatlashtirish hisoblanadi. Davlat boshqaruvi tizimida “Elektron hukumat” tizimini yaratish bo‘yicha ulkan dasturlar ilgari surilib, hayotga tadbiq etilmoqda. Ushbu dasturlarning muhim tarkibiy qismi atamalar va tushunchalarni har xil tushunishlarni bartaraf etish uchun ularning bir ma’noli izohlarini, birlashgan hujjatlarni yaratishdir.

To‘g‘ri yozishni tekshiruvchi tizimlar nisbatan kam sonli tabiiy tillar uchun ishlab chiqilgan. Amaliyotda bunday tizimlarning qo‘llanilishi xatolarni tezkor aniqlash va to‘g‘rilashga yordam beradi. Nokorrekt masalalarni yechishda regulyarizatsiyalash yondashuvi mavjud bo‘lib, optimallashtirish masalasi yetarlicha aniqlanmagan holda, asosiy kriteriyaga yechilayotgan masalaning o‘ziga xosligi va predmet sohaning bilimlarini hisobga oladigan qo‘shimcha kriteriya – regulyarizator qo‘shiladi. Matnlarga avtomatlashtirilgan ishlov berishning amaliy masalalarida, ular zaif shakllanganligi sababli, yechimlari ko‘p bo‘lishi mumkin bo‘lgan holatlarda qo‘shimcha kriteriya va chegaralar ishlab chiqishga to‘g‘ri keladi.

Anglash usullaridan foydalanishning nazariyasi va amaliyotining rivojlanishida chet el va yurtimiz olimlari ulkan hissalarini qo‘shishgan. Chet el olimlari orasida Yu.I. Juravlev, N.G.Zagoruyko, V.A.Dyuk, K.V.Vorontsov,

M.Livenshteynlarning, yurtimiz olimlari ichida M.M.Kamilov, T.F.Bekmuratov, F.T.Adilova, G.R.Matlatipov, M.Musayevlarning tadqiqotlarini alohida qayd etish mumkin.

Dissertasiya tadqiqotining dissertasiya bajarilgan oliy ta'lim muassasining ilmiy tadqiqot ishlari bilan bog'liqligi. Dissertasiya tadqiqoti O'zbekiston Milliy universitetining ilmiy tadqiqot ishlari rejasiga muvofiq F-4-64 "Berilganlarni intellektual tahlilida umumlashgan baholarni hisoblash va obyektlarning individual metrikasiga asoslangan usullarni ishlab chiqish va asoslash" (2011-2016) ilmiy tadqiqot loyihasi doirasida bajarilgan.

Tadqiqotning maqsadi tematik modellashtirish masalasini yechish uchun berilganlarning intellektual tahlili usullari va kriteriyalarini ishlab chiqishdan iborat.

Tadqiqotning vazifalari:

tabiiy tildagi hujjatlar kolleksiyasi uchun lug'at yaratish jarayonini avtomatlashtirish;

hujjatlar sinflarining kompaktligini bog'langanlik munosabatlari bo'yicha baholash;

kolleksiyadagi hujjatlarning semantik bog'langanlik darajasini aniqlash;

hujjatlar sinflarining juftligi bo'yicha so'zlardan sun'iy qoliplarni ajratib olish;

hujjatlarning mazmuniy haqqoniylik parametrini hisoblash metodini ishlab chiqish.

Tadqiqot obyekti tabiiy tillardagi hujjatlarning semantikasini tahlil qilish texnologiyalarini ishlab chiqish va asoslashdan iborat.

Tadqiqot predmeti. Tematik modellashtirish masalalarida hujjatlar kolleksiyasining tahlili uchun sinflashning metrik metodlari.

Tadqiqot usullari. Tadqiqot ishida diskret matematika, sun'iy intellekt, matematik tahlil va algoritmik tillar orqali dasturlash asosida anglashning optimal algoritmlarini izlash usullaridan foydalanilgan.

Tadqiqotning ilmiy yangiligi quyidagilardan iborat:

hujjatlar tavsifidagi munosabatlarining miqdoriy bahosi asosida ularning semantik bog'langanlik koeffitsiyentini hisoblash usuli taklif qilingan va asoslangan. Ilmiy uslubdagi hujjatlarda koeffitsiyentning eng yuqori qiymati matematika va fizika fanlari bo'yicha olingan;

xom alomatlarning o'zaro kesishmaydigan guruhlari bo'yicha umumlashgan baholarni hisoblash metodi orqali shakllantirilgan latent alomatlar yordamida hujjatlar sinflari juftligi bo'yicha so'zlardan sun'iy qoliplar ajratib olingan;

hujjatlar kolleksiyasining mazmunan haqqoniylik ko'rsatkichini hisoblash metodologiyasi ishlab chiqilgan. Metodikada, mavzularning optimal sonini aniqlash uchun klaster tahlil natijalari bo'yicha kriteriyalardan foydalanilgan;

ikkita hujjatlar to'plamidagi berilganlar uchun terminlarning umumiy lug'atini shakllantirish metodi ishlab chiqilgan. Lug'at uchun terminlarni tanlashda ularning sinfga tegishlilik funksiyasining qiymati bo'yicha hisoblanuvchi turg'unlik ko'rsatgichidan foydalanilgan.

Tadqiqotning amaliy natijasi hujjatlarni mavzu bo'yicha, ularning semantik bog'langanligini hisobga olgan holda avtomatik sinflash imkoniyatida aks etadi.

Tadqiqot natijalarining ishonchliligi. Olingan natijalarning ishonchliligi, semantik hisoblash metodlaridan foydalangan holda tematik modellashtirish masalasining yechimi bo‘yicha barcha hisoblash eksperimentlarining predmetga yo‘naltirilgan sohalar uchun tematik sinflashning adekvatligini tasdiqlashi bilan asoslanadi.

Tadqiqot natijalarining ilmiy va amaliy ahamiyati. Tadqiqot natijalarining ilmiy ahamiyati hujjatlar sinflari va guruhlarining bog‘langanligi hamda semantik bog‘lanish koyeffitsiyentlarini hisoblash usullarini ishlab chiqish va asoslashda, hujjatlar to‘plami uchun umumiy lug‘atlarni shakllantirishda, hujjatlarning mazmuniy haqqoniyligini hisoblash algoritmida ko‘rinadi.

Amaliy ahamiyati turli predmet sohalardagi so‘zlarning ko‘p ma’nolilik masalasini yechish va yangi bilimlarni olishda foydalaniladigan, semantik tahlil metodining algoritmlarini amalga oshirish uchun dasturiy ta’minotni ishlab chiqishda ko‘rinadi.

Tadqiqot natijalarining joriy qilinishi. Hujjatlar tavsifidagi munosabatlarining miqdoriy bahosi asosida ularning semantik bog‘langanlik koeffitsiyentini hamda hujjatlar kolleksiyasining mazmuniy haqqoniylik ko‘rsatkichini hisoblash orqali hujjatlar to‘plamidagi berilganlar uchun terminlarning umumiy lug‘atini shakllantirish metodi ishlab chiqish orqali olingan ilmiy natijalar asosida:

hujjatlar to‘plamlari orasidagi semantik bog‘lanishni hisoblash metodini yaratish bo‘yicha dissertatsiya ishida olingan natijalardan JHBL-13 raqamli “O‘zbekistonda “Baxtiyor oila” veb portalini yaratish” grant loyihasida matnli ma’lumotlarni xronologik tahlil qilishda foydalanilgan (“Mahalla va oila” ilmiy-tadqiqot institutining 2021-yil 25-iyundagi 10-sonli ma’lumotnomasi). Ilmiy natijalarning qo‘llanilishi matnli hujjatlarni klassifikatsiya qilish uchun alomatlar fazosini tashkil qiluvchi lug‘atlarni shakllantirish imkonini bergan;

hujjatlar to‘plami bo‘yicha umumiy lug‘atni shakllantirish va hujjatlar tavsifi munosabatlarini miqdoriy baholash bo‘yicha olingan natijalar “UNICON.UZ” Fan-texnika va marketing ilmiy tadqiqotlar markazi Davlat unitar korxonasida o‘zbek tilidagi terminlar glossariylarini yaratish hamda elektron hujjat aylanish tizimlarida qo‘llanilgan (“UNICON.UZ” DUK 2023-yil 20-fevraldagi 1-1/300-sonli ma’lumotnomasi). Tematik modellarni ishlab chiqish uchun dasturiy ta’minoti ilmiy terminlar lug‘ati va ular asosida elektron hujjatlar yaratish jarayonini avtomatlashtirish hisobiga qaror qabul qilishning tezkorligi va samaradorligini 40 % oshirishga yordam bergan.

Tadqiqot natijalarining aprobatsiyasi. Mazkur tadqiqot natijalari 6 ta, jumladan, 4 ta xalqaro va 2 ta respublika ilmiy-amaliy anjumanlarida muhokamadan o‘tkazilgan.

Tadqiqot natijalarining e’lon qilinganligi. Tadqiqot mavzusi bo‘yicha jami 16 ta ilmiy ish chop etilgan, shulardan, O‘zbekiston Respublikasi Oliy Attestasiya komissiyasining doktorlik dissertasiyalari asosiy natijalarini chop etish tavsiya etilgan ilmiy nashrlarda 6 ta maqola, jumladan, 3 tasi xorijiy va 3 tasi respublika jurnallarida nashr etilgan. Shuningdek, yaratilgan kompyuter dasturiy mahsulotlari uchun 3 ta mualliflik guvohnomasi olingan.

Dissertasiyaning tuzilishi va hajmi. Dissertatsiya kirish qismi, uchta bob, xulosa, foydalanilgan adabiyotlar ro‘yxati va ilovalardan tashkil topgan. Dissertasiyaning hajmi 110 betdan iborat.

DISSERTATSIYANING ASOSIY MAZMUNI

Kirish qismida dissertatsiya mavzusining dolzarbligi va zarurati, tadqiqotning respublika fan va texnologiyalarni rivojlantirishning ustuvor yo‘nalishlariga mos kelishi asoslangan, dissertasiya mavzusi bo‘yicha chet eldagi ilmiy tadqiqotlarning qisqacha ma’lumoti va muammoning o‘rganilganlik darajasi keltirilgan, tadqiqotning maqsad, vazifalari shakllantirilgan, uning obyekti va predmeti ko‘rsatilgan, tadqiqotning amaliy natijalari va ilmiy yangiliklari bayon qilingan, olingan natijalarning nazariy va amaliy ahamiyati ochib berilgan, tadqiqot natijalarining qo‘llanilishi, dissertasiya tuzilishi va nashr qilingan ilmiy ishlar to‘g‘risida ma’lumotlar keltirilgan.

“Tabiiy tilning berilganlar modellari” nomli I bobda tematik modellashtirishning o‘ziga xosliklari qaralgan. Shu maqsadda 1.1-§ da tabiiy tildagi matnlarning semantik tahlilidagi terminologiyasi keltirilgan. So‘zlar, tushunchalar, tabiiy til qoliplarining semantikasi kontekstdan kelib chiqib aniqlanganligi sababli, mazmunini boshqa so‘zlar yordamida tushuntirish talab etiladi. Buning uchun ontologiyaning har xil ko‘rinishlaridan foydalaniladi. Ushbu ko‘rinishlar ontologiyaning quyidagi maqsadlarga yo‘nalgan bo‘limlarini o‘z ichiga oladi:

– lug‘at – bir qiymatli terminlar ro‘yxati;

– glossariy – ko‘p ma’noli terminlarning ma’nolarini hisobga oluvchi lug‘at;

– tezaurus – berilgan semantik bog‘lanish tizimiga ega glossariy.

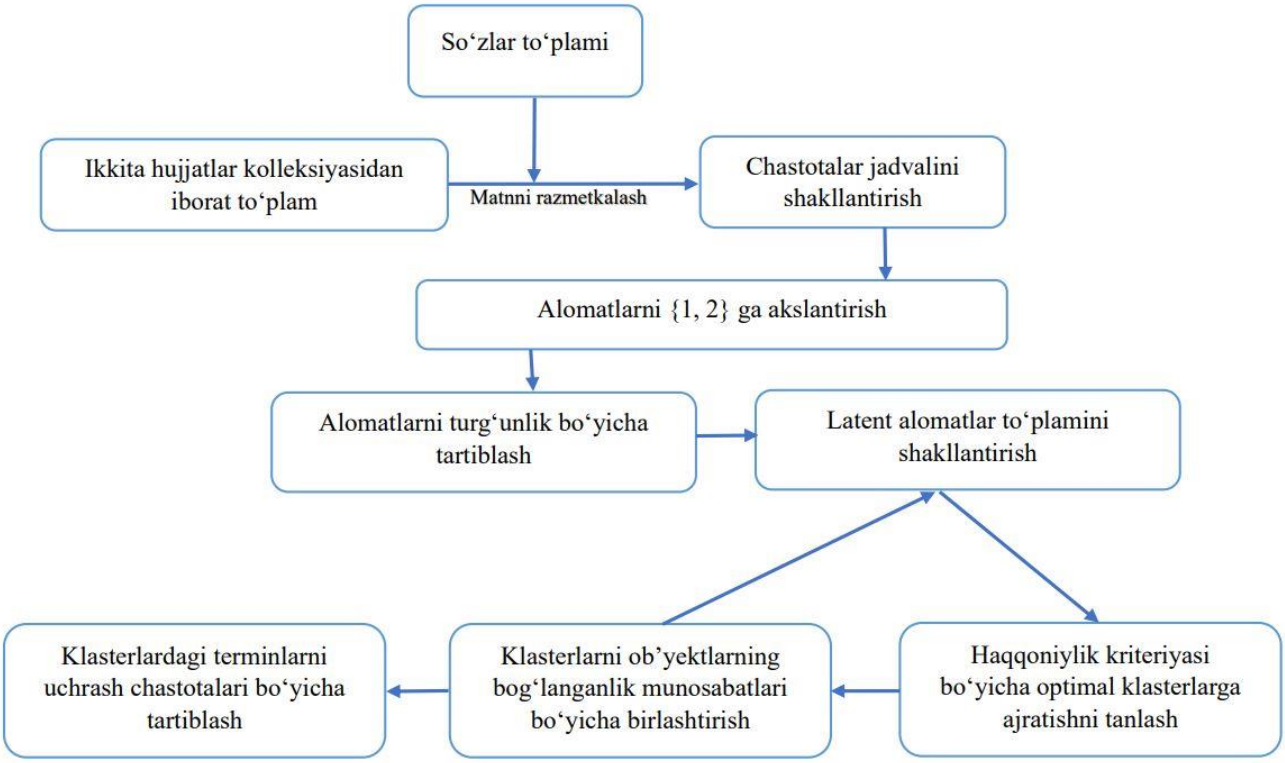
Tematik modellashtirishda eng ko‘p ishlatiladigan atamalar tavsiflangan. Predmet sohaga yo‘naltirilgan lug‘atlar uchun ontologiyada qo‘llaniladigan tushunchalarga urg‘u berilgan. Lug‘atlarni shakllantirishning ushbu tadqiqot ishida taklif etilayotgan funksional sxemasi 1-rasmda ko‘rsatilgan.

1.2-§da sinflash algoritmlari yordamida hujjatlarning klaster tahlili ko‘rib chiqiladi. Matnli hujjatlar to‘plamini klasterlash, ularning ma’no jihatdan yaqin bo‘lgan kichik to‘plamlarga bo‘linishini anglatadi.

Hujjatlarni tavsiflash uchun fazoni tanlash maqsadida alomatlarni klasterlash ko‘rib chiqiladi. Bu o‘rinda “o‘lcham la’natisi” sifatda ma’lum bo‘lgan BigData muammosi diqqatga olingan. Klasterlash jarayoni obyektlarning umumlashgan bahosini hisoblash metodi bilan xom alomatlarning o‘zaro kesishmaydigan guruhlari bo‘yicha latent alomatlarni shakllantirish orqali amalga oshiriladi. Tadqiqot sxemasi 2-rasmda ko‘rsatilgan.

An’anaviy usullar ro‘yxati gipersharlar ko‘rinishidagi mantiqiy qonuniyatga asoslangan hujjatlarning juft bog‘langanlik munosabatidan foydalanadigan guruhlash algoritmlari bilan to‘ldirilgan. 1.3-§da tematik hujjatlar bazasini shakllantirishdan maqsad-tabiiy tilning qonuniyatlari va xususiyatlarining tahlili ekanligi qayd qilingan. Bitta tilda yozilgan hujjatlar to‘plami va hujjatlar to‘plamidagi ushbu matnlar haqidagi qo‘shimcha ma’lumotlar korpusni tashkil qiladi. Korpus tabiiy til modellarini yaratish va tahlil qilish uchun muhim ahamiyatga ega. Korpus nutqning turli xil uslublarida (publitsistik, ilmiy, rasmiy, badiiy) yozilgan hujjatlarni o‘z ichiga oladi. So‘zlardan ko‘ra tushunchalar ko‘proq

bo‘lganligi sababli, so‘zlar turli mavzularda o‘z ma’nosi bilan ishlatilishi mumkin. Bunday hollarda to‘g‘ri ma’noni tanlashga so‘zlar orasidagi semantik munosabatlar yordam beradi. Bundan tashqari, so‘z boshqa so‘z bilan bog‘lanib kelsa, ko‘proq ma’no berishi mumkin. Ushbu maqsadlarga leksemalar to‘plami sifatida ma’lum bir fan sohasining asosiy semantik munosabatlari yig‘indisini aks ettiruvchi tematik (semantik) tamoyil bo‘yicha tashkil etilgan fan sohasiga yo‘naltirilgan lug‘at xizmat qiladi.



1-rasm. Lug‘atlarni shakllantirish jarayoni.

Dissertatsiyaning “Ilmiy uslubda yozilgan hujjatlarning to‘plamlari bo‘yicha tematik modellashtirish” nomli II bobida ilmiy uslubda yozilgan hujjatlar tahlilining o‘ziga xosliklari tahlil qilingan. 2.1-§da ilmiy uslub O‘zbekiston Respublikasi OAK bazasidagi ilmiy dissertatsiyalar avtoreferatlarini tahlil qilish orqali taqdim etiladi.

Mashinada ishlov berish uchun matnli hujjatlar to‘plami shaklidagi ochiq ma’lumotlar manbalari nisbatan kamligi sababli o‘zbek tili, ilmiy tadqiqot natijalarini taqdim etish nuqtayi nazaridan kam tarqalgan tilga misol bo‘ladi. Bu hujjatlarga qo‘yiladigan talablardan biri shundaki, ular o‘zbek tili grammatikasi qoidalariga muvofiq yozilgan bo‘lishi shart. Dastlabki ishlov berish ma’nosi, asosan, barcha so‘zlarni lemmatizatsiya yoki stemming o‘tkazish orqali normal shaklga keltirishga mos keladi. Avtoreferatlar to‘plamlari bilan ishlashga undovchi sabablar quyidagilar edi.

1. O‘zbek tilining ommaviy foydalanish uchun ochiq bo‘lgan va rasman tasdiqlangan korpusi mavjud emas. Mashinali o‘rgatishda foydalanish uchun

so‘zlarga dastlabki ishlov berishning normal shaklini tanlash masalasi ochiqligicha qolmoqda.

2. Avtoreferatlar mazmuni dissertatsiya materiali bayonining o‘xshash tuzilishiga ega bo‘lgan uchta o‘zbek, rus yoki ingliz tillarida beriladi. Rus va ingliz tillari korpusining mavjudligi hamda ular asosida matnni tahlil qilish usullari o‘zbek tilidagi avtoreferatlar mazmunini qiyosiy tahlil qilish imkonini beradi.

3. Tematik modellashtirishda mavzular soni va ular mazmunining o‘zgarish sabablarini tushuntirish uchun turli ilmiy sohalar orasidagi aloqalarni izlash.

4. Dissertatsiyalar mazmunini baholash uchun ekspertlar, taqrizchilarning bazasini shakllantirish.

Tajriba uchun 1634 ta hujjatdan foydalanilgan bo‘lib, ularning fan sohalari bo‘yicha taqsimoti 1-jadvalda keltirilgan.

1-jadval.

Fan sohalari bo‘yicha hujjatlarning taqsimlanishi

Fan sohasi	Hujjatlar soni	Fan sohasi	Hujjatlar soni
Biologiya	109	Huquqshunoslik	83
Fizika	162	Iqtisodiyot	189
Geografiya	28	Kimyo	120
Geologiya	44	Madaniyatshunoslik	11
Matematika	95	Texnika	380
Pedagogika	137	Tibbiyot	277

Tematik hujjatlarning bog‘langanligi va kompaktlik o‘lchamini hisoblash 2.2-§da masalaning quyidagicha qo‘yilishi asosida amalga oshiriladi.

$E_0=\{S_1,\ldots,S_m\}$ obyektlar to‘plami l ($l\geq 2$) ta o‘zaro kesishmaydigan $K_1,\ldots,K_l$ qism to‘plamlarga (sinflarga) ajralgan holda berilgan bo‘lsin. Obyektlarning tavsifi n ta $X(n)=(x_1,\ldots,x_n)$ miqdoriy alomatlar yordamida amalga oshiriladi. E_0 obyektlarning tavsiflari to‘plamida $\rho(x,y)$ metrika berilgan va $\mu(E_0,\rho,R^n)$ sinflarning kompaktlik o‘lchami aniqlangan bo‘lsin. Talab qilinadi:

– $X(d)\subset X(n)$, $d\leq n$ shovqin ob’yektlarsiz to‘plamni aniqlash;

– obyektlar tavsiflanishini R^d fazodan R^k o‘lchamli fazoga $R^d\rightarrow R^k$, $k<d$ o‘girishni amalga oshirish;

– $\mu(E_0,\rho,R^n)$ bo‘yicha R^d fazoning optimal o‘lchamini aniqlash $d=\arg\max\limits_{t\in\{2,\ldots,n\}}\mu(E_0,\rho,R^t)$.

K_t, $t=1,\ldots,l$ sinfning kompaktlik o‘lchamining qiymati (2-jadvalga qarang) obyektlarni o‘zaro kesishmaydigan $G_{t1},\ldots,G_{t\beta(t)}$, $\beta(t)\geq 1$ guruhlarga bo‘lish orqali quyidagicha

$$\Theta_t=\frac{\sum_{i=1}^{\beta(t)}|G_{ti}|^2}{|K_t|^2} \qquad (1)$$

va butun tanlanmaniki esa quyidagicha

$$\Theta=\frac{\sum_{i=1}^{l}|K_i|\Theta_i}{m} \qquad (2)$$

hisoblanadi.

2-jadval.

Fan sohalari bo‘yicha hujjatlarning kompaktlik o‘lchamining qiymatlari

Fan sohasi	Kompaktlik o‘lchami	
	fan sohasi bo‘yicha	tanlanma bo‘yicha
Biologiya	0.4370	0.5686
Fizika	0.0515	
Geografiya	0.2976	
Geologiya	0.5103	
Huquqshunoslik	0.9294	
Iqtisodiyot	0.7829	
Kimyo	0.3655	
Madaniyatshunoslik	0.6859	
Matematika	0.0145	
Pedagogika	0.9426	
Texnika	0.5138	
Tibbiyot	0.8678	

Tanlanma bo‘yicha kompaktlik (2) 0.5686 ga teng. Turli fan sohalari terminologiyasining o‘xshashligini baholash uchun 2.3-§da quyidagicha masala qo‘yilgan.

$E_0=\{S_1,\ldots,S_m\}$ ob’yektlar to‘plami l ($l\geq 2$) ta o‘zaro kesishmaydigan $K_1,\ldots,K_l$, $E_0=\bigcup_{i=1}^{l} K_i$ to‘plam ostilarga (sinflarga) ajralgan holda berilgan bo‘lsin. Obyektlarning tavsifi n ta $X(n)=(x_1,\ldots,x_n)$ miqdoriy alomatlar yordamida amalga oshiriladi. E_0 obyektlarining tavsiflari to‘plamida $\rho(x,y)$ metrika berilgan va $\Theta(K_i,E_0,\rho)$, $\Theta(K_j,E_0,\rho)$ kompaktlik o‘lchami hamda $K_i\cup K_j$ $\Theta(K_i\cup K_j,E_0,\rho)$, $i,j\in\{1,\ldots,l\}$ sinflar birlashmalari aniqlangan bo‘lsin. $\Theta(K_i,E_0,\rho)$, $\Theta(K_j,E_0,\rho)$, $\Theta(K_i\cup K_j,E_0,\rho)$ bo‘yicha sinflarning bog‘langanlik o‘lchami va o‘xshashlikning juftlik o‘lchamini aniqlash talab qilinadi.

$\theta_i,\theta_j,\theta_{ij}$ - qiymatlar mos ravishda K_i, K_j va $\bigcup_{d=1}^{l} K_d$, $l>2$ da $K_i\cup K_j$, $i,j\in\{1,\ldots,l\}$ lardagi hujjatlarning (1), (2) ga muvofiq aniqlanuvchi kompaktlik o‘lchamlari bo‘lsin. U holda $\bigcup_{d=1}^{l} K_d$ dagi $\varphi(K_i, K_j)$ hujjatlar o‘xshashligining juftlik o‘lchami quyidagicha hisoblanadi:

$$\varphi(K_i, K_j) = 1 - \frac{\theta_i}{\theta_{ij}}\frac{\theta_j}{\theta_{ij}}. \quad (3)$$

Obyektlar sinflarining juftliklari to‘plamini (3) ga muvofiq ajratib olamiz

$$\Omega=\{(K_i,K_j)|\varphi(K_i,K_j)>0,\ i,j\in\{1,\ldots,l\}\}. \quad (4)$$

va ushbu

$$R(K_i,K_j) = \frac{\max(\theta_i,\theta_j)}{\theta_{ij}} \quad (5)$$

munosabatni hujjatlarning semantik bog‘langanligi sifatida aniqlaymiz.

3-jadval.

Ilmiy sohalar juftliklari bo‘yicha o‘xshashlik o‘lchamlari

Fan sohasi bo‘yicha hujjatlar	Kompaktlik o‘lchami		$1-\frac{\theta_i}{\theta_{ij}}\frac{\theta_j}{\theta_{ij}}$
	θ_i, θ_j	θ_{ij}	
Matematika, fizika	0.0145 0.0804	0.3357	0.9896
Matematika, tibbiyot	0.0145 0.8745	0.5237	0.9537
Matematika, texnika	0.0145 0.5176	0.5253	0.9440
Matematika, iqtisodiyot	0.0145 0.7829	0.3484	0.9063
Matematika, pedagogika	0.0145 0.9426	0.3361	0.8788
…			
Geografiya, fizika	0.2857 0.0804	0.0647	-4.4911
Geologiya, fizika	0.5103 0.0804	0.0749	-6.3228
Geografiya, matematika	0.2857 0.0145	0.0235	-6.5322
Madaniyatshunoslik, fizika	0.6859 0.0804	0.0733	-9.2629
Madaniyatshunoslik, matematika	0.6859 0.0145	0.019	-26.4481

Ushbu fanlar bo‘yicha hujjatlarni bitta sinfga birlashtirganda (3-jadvalga qarang) kompaktlik o‘lchamining 0.3357ga teng bo‘lgan qiymati olingan. Bu natija fizika-matematika fanlari bo‘yicha ilmiy darajalarni berish uchun ilmiy tasnifni tanlashda o‘z aksini topgan holda, ushbu fanlarda o‘zaro bog‘liq atamalar mavjudligini ko‘rsatadi.

(3), (4) asosida shakllantirilgan va (5) ga ko‘ra tartiblangan sohalar juftliklari bo‘yicha hujjatlar to‘plamlarining ketma-ketligi 4-jadvalda keltirilgan.

4-jadval.

Hujjatlarning semantik bog‘langanligi

Fan sohalari	Munosabat (5)	Fan sohalari	Munosabat (5)
Matematika, fizika	0.2397	Biologiya, tibbiyot	0.9672
Kimyo, texnika	0.7259	Huquqshunoslik, pedagogika	0.9869
Geografiya, geologiya	0.8967	Biologiya, geologiya	0.9933
Geografiya, texnika	0.9122	Kimyo, biologiya	0.9945
Geologiya, texnika	0.9298	Geografiya, iqtisod	1.0084

“Hujjatlarning mazmuniy haqqoniyligi” nomli III bobda metrik algoritmlar yordamida amalga oshiriladigan berilganlarning klaster tahlili uchun hujjatlar

mazmunining haqqoniylik ko‘rsatkichlarini hisoblash muhokama qilinadi. Big Data bilan bog‘liq vazifalardan biri hujjatlarning umumiy lug‘atidagi terminlarni tanlashdir.

3.1-§da matematik lingvistika masalalarida tabiiy til atamalaridan sun’iy qoliplarni ishlab chiqish keltirilgan. Qoliplar hujjatlar sinflari juftligiga nisbatan latent yoki meta-alomatlar sifatida shakllantiriladi.

Masalaning qo‘yilishi. O‘zaro kesishmaydigan 2 ta K_1, K_2 sinflarga bo‘lingan $E_0=\{S_1,\ldots,S_m\}$ obyektlar tanlanmasi berilgan. E_0 dagi obyektlar tavsifi n ta $X(n)=(x_1,\ldots,x_n)$ miqdoriy alomatlar yordamida ifodalanadi. Ushbu jarayonlar aniqlangan:

– K_1,K_2, $t\geq 1$, $\forall c\in\{1,\ldots,t\}$ $n_c>1$, $n_1+\ldots+n_t\leq n$ bo‘yicha $X(n)$ ni o‘zaro kesishmaydigan $X(n_1),\ldots, X(n_t)$ to‘plam ostilariga ajratish va ular bo‘yicha $Y(t)=(y_1,\ldots,y_t)$ latent alomatlar to‘plamlarini hisoblash;

– E_0 obyektlarini berilgan sondagi o‘zaro kesishmaydigan $G_1,\ldots,G_p$ klasterlarga (guruhlarga) ajratish;

– berilgan p sondagi guruhlarga ajralish sifatining $F(p)$ bahosini hisoblash;

Talab qilinadi:

– E_0 ni p guruhlarga ajratish uchun $Y(\sigma)\subset Y(t)$ latent alomatlar to‘plamini ajratish;

– $\mu=\arg\max_p F(p)$ guruhlar sonini aniqlash;

– $G_1,\ldots,G_\mu$ guruhlarning har biri bo‘yicha $\Omega=\{X(n_c)|y_{n_c}\in Y(\sigma)\}$ to‘plamdagi so‘zlar rangini hisoblash;

– ikkita hujjatlar sinfi vakillaridagi eng ko‘p ishlatilgan so‘zlarning rangini aniqlash.

$K_1 \cup K_2$ ob’yektlari tavsifidagi $x_c\in X(n)$ alomat qiymatlari uchun kamaymaydigan tartibdagi ketma-ketligi qurilgan bo‘lsin

$$R_1,\ldots,r_j,\ldots,r_m,\ m=|K_1\cup K_2|. \qquad (6)$$

(6) ni ikkita $[c_1, c_2]$ va $(c_2, c_3]$ intervallarga ajratamizki, uning chegaralari quyidagi kriteriyaga muvofiq aniqlanadi

$$\left(\frac{\sum_{p=1}^{2}\sum_{i=1}^{2}u_i^p\left(m-|K_i|-\sum_{j=1}^{2}u_j^p-u_i^p\right)}{2|K_1||K_2|}\right)\left(\frac{\sum_{p=1}^{2}\sum_{i=1}^{2}u_i^p\left(u_i^p-1\right)}{|K_1|(|K_1|-1)+|K_2|(|K_2|-1)}\right)\rightarrow\max_{c_1<c_2<c_3}, \qquad (7)$$

bu yerda u_i^p - p intervaldagi K_i sinf obyektlarining soni.

(6) dagi ushbu

$$\left|\frac{d_{t,c}(u,v)}{|K_t|}-\frac{d_{3-t,c}(u,v)}{|K_{3-t}|}\right|\rightarrow\max \qquad (8)$$

kriteriya bo‘yicha aniqlanadigan $[r_u;r_v]^i$ intervalning chegaralaridagi qiymatlar uchun interval nomeri unga mos nominal alomatning gradatsiyasi sifatida qaraladi.

$D_{tc}(u,v)$, $d_{3\text{-}t,c}(u,v)$ – $[r_u;r_v]^i$, $i\in\{1,\ldots,p_c\}$ intervaldagi K_t, $K_{3\text{-}t}$ sinf vakillarining soni bo‘lsin. $[r_u;r_v]^\mu$ ($\mu\in\{1,\ldots,p_c\}$ gradatsiyalar) interval bo‘yicha K_1 sinfga tegishlilik funksiyasining qiymati $f_c(\mu)$ quyidagicha hisoblanadi:

$$f_c(\mu)=\frac{\frac{d_{1c}(\mu)}{|K_1|}}{\frac{d_{1c}(\mu)}{|K_1|}+\frac{d_{2c}(\mu)}{|K_2|}}. \qquad (9)$$

E_0 obyektlarning umumlashgan bahosini hisoblash uchun alomatlarning gradatsiyalari ulushidan foydalaniladi. $X_c \in X(n)$ alomatning $j \in \{1,2\}$ gradatsiyalari ulushi quyidagicha aniqlanadi:

$$\eta_c(j) = w_c\left(\frac{\alpha_{cj}^1}{|K_1|} - \frac{\alpha_{cj}^2}{|K_2|}\right), \quad (10)$$

bu yerda α_{cj}^1, α_{cj}^2 – mos ravishda K_1 va K_2 sinflardagi x_c alomatning j gradatsiyasining qiymatlari soni, w_c – x_c alomatning sinflararo o‘xshashlik va farqlanishlar ko‘paytmasi sifatida aniqlanuvchi vazni. $S_r \in E_0$ obyektning umumlashgan bahosi $X(n)$ to‘plamdagi $S_r = \{a_{ri}\}_{i \in X(n)}$ o‘lchamlarining nominal shkaladagi tavsifi bo‘yicha va (10) ulush bilan quyidagicha hisoblanadi:

$$R(S_r) = \sum_{i \in X(n)} \eta_i(a_{ri}). \quad (11)$$

$S_r \in K_1 \cup K_2$ obyekt tavsifida $X(n)$ dagi alomatlarning dastlabki qiymatlari (9) bo‘yicha $S_r = \{b_{ri}\}_{i \in X(n)}$ tegishlilik funksiyasining qiymatlari bilan almashtirilgan bo‘lsin. $x_c \in X(n)$ alomatning turg‘unligi quyidagicha hisoblanadi:

$$\varphi(c) = \frac{1}{m}\sum_{r=1}^{m} \begin{cases} b_{rc}, b_{rc} > 0.5 \\ 1 - b_{rc}, b_{rc} < 0.5 \end{cases}. \quad (12)$$

So‘zlarning (12) bo‘yicha tartibi umumiy holda hujjatlar to‘plami tavsifi uchun fazoning o‘lchamiga bog‘liq emas.

Mazmunning haqqoniylik koeffitsiyentini hisoblashning ma’nosi latent alomatlar fazosida hujjatlarning mavzularga bo‘linishini aniqlaydigan klasterlarning optimal sonini izlashdir. Latent fazo va u asosida umumiy lug‘atni shakllantirish uchun quyidagilar taklif etiladi:

– D dagi indekslarga ega alomatlarni ularning (12) turg‘unligi qiymatlariga ko‘ra o‘sish bo‘yicha tartiblash

$$\varphi(\varepsilon_1), \ldots, \varphi(\varepsilon_j), \ldots, \varphi(\varepsilon_{dim}), \varepsilon_j \in D, dim = |D| \quad (13)$$

– (13) bo‘yicha $Y(t)=(y_1,\ldots,y_t)$ latent alomatlar to‘plamini shakllantirish;

– umumiy lug‘at sifatida D bo‘yicha $Y(\sigma)$ xom alomatlar bilan bog‘langan $Y(\sigma) \subset Y(t)$, $\sigma \leq t$ va to‘plamni ajratish.

(13) dagi umumiy lug‘at uchun so‘zlar soni, umuman olganda, erkin parametrdir. Ushbu parametr evristik yo‘l bilan yoki latent alomatlar to‘plamini shakllantirishda iyerarxik aglomerativ guruhlash natijalariga asosan berilishi mumkin.

Quyidagicha belgilashlarni kiritamiz: *TYPLAM* – iyerarxik guruhlash algoritmi yordamida guruh tarkibiga qo‘shilgan xom alomatlar indekslarining to‘plami, *lugat* – umumiy lug‘utdagi so‘zlar soniga chegara, *guruh* – latent alomatlar soni. Algoritmning qadamlar bo‘yicha amalga oshirilishi quyidagicha bo‘ladi.

1 Qadam. $j=0$. *guruh*=0.

2 Qadam. Hisoblash $j=j+1$. *crit*=10. $u=\varepsilon_j$. *TYPLAM* = {u}. *guruh*=*guruh*+1.

Qadam $t \in \{1,\ldots,m\}$ $R(S_t)=\eta_u(a_{tu})$ bo‘yicha. Sikl tugashi;

3 Qadam. $u=\varepsilon_{j+1}$. Sikl $t \in \{1, \ldots, m\}$ $b_t = R(S_t) + \eta_u(a_{tu})$ bo‘yicha. Sikl tugashi;

$M_1 = \sum_{S_t \in K_1} b_t$. $M_2 = \sum_{S_t \in K_2} b_t$. $M_1=M_1/|K_1|$. $M_2=M_2/|K_2|$. $\theta=0$. $\gamma=0$. Sikl $t \in \{1, \ldots, m\}$ bo‘yicha. Agar $S_t \in K_1$, u holda $\theta=\theta+|b_t-M_1|$, $\gamma=\gamma+|b_t-M_2|$. Aks holda $\theta=\theta+|b_t-M_2|$, $\gamma=\gamma+|b_t-M_1|$. Sikl tugashi;

4 Qadam. Agar $\theta/\gamma < crit$, u holda $crit = \theta/\gamma$, $TYPLAM = TYPLAM \cup \{u\}$, $j = j+1$, o‘tish 3.

5 Qadam. Chiqarish $\{R(S_t)\}_{t \in \{1, ..., m\}}$, *TYPLAM*.

6 Qadam. Agar $j<lugat$, u holda o‘tish 2; Aks holda *guruh* chiqarish.

7 Qadam. Tugatish.

Latent alomatlarni shakllantirishni ifodasi 5-jadvalda keltirilgan. K_1 sinf fizika va matematika, K_2 qolgan barcha 10 ta fan sohalarining hujjatlari bilan ifodalangan.

5-jadval.

Latent alomatlar va ularning xom alomatlardagi tarkibi

№	Alomatlar soni	Alomatlar	(7) kriteriya qiymati
1	267	179, 55, 171, … , 414, 297, 299	0.6360
2	87	423, 66, 450, … , 186, 36, 265	0.5061
3	66	70, 67, 192, …. , 290, 15, 21	0.3917
4	10	393, 61, 75, 253, 163, 405, 53, 24, 44, 490	0.2793
5	23	322, 408, 295, … , 311, 364, 316	0.3493
6	1	89	0.2442
7	16	139, 189, 166, … , 305, 199, 304	0.3492
8	8	50, 158, 223, 195, 427, 302, 433, 369	0.2372
9	2	98, 197	0.2127
10	1	177	0.1852
11	2	57, 279	0.2344
12	2	120, 135	0.2460

Y(*guruh*) latent alomatlar to‘plami bo‘yicha o‘zaro kesishmaydigan guruhlarga bo‘linishi $G_1,...,G_h$, $i=1, ...,h$, $h \geq 2$ bo‘lsin. Har bir G_i guruh uchun obyektlarning G_i bo‘yicha K_1 sinfga tegishlilik funksiyasining qiymatini $\lambda_i(K_1)=d_{i1}/|G_i|$ kabi aniqlanadi, bu yerda d_{i1}– G_i dagi K_1 sinf obyektlarining soni. E_0 dagi hujjatlarning *h* guruhlarga bo‘lingandagi mazmuniy haqqoniyligi

$$F\big(h, Y(guruh)\big) = \frac{1}{m}\sum_{j=1}^{h} \begin{cases} |G_j|\lambda_j(K_1), \lambda_j(K_1) > 0.5; \\ |G_j|(1-\lambda_j(K_1)), \lambda_j(K_1) < 0.5. \end{cases} \quad (14)$$

kabi hisoblanadi.

Umumiy lug‘at yordamida mazmunning haqqoniyligini hisoblash uchun (14) dan foydalanishning maqsadga muvofiqligi quyidagicha tekshirilishi mumkin.

Obyektlarning $G_1, ...,G_h$ guruhlarga bo‘linishi va ularning har birida K_1 yoki K_2 sinflardan qay birining vakillari ustunlik qilishi ma’lum deb hisoblangan. Mos ravishda K_1 va K_2 sinf vakillari ustunlik qiluvchi guruhlardan iborat ikkita o‘zaro kesishmaydigan Q_1 va Q_2 to‘plam ostilarning birlashmasi $T = Q_1 \cup Q_2$ sifatida ifodalangan $T=(M_1, ...,M_h)$ orqali guruh (obyektlar) markazlarining to‘plamini belgilaylik. Guruh markazlari to‘plam ostilarining optimal sonini (keyingi o‘rinlarda hujjat mavzulari deb yuritiladi) aniqlash uchun o‘zaro kesishmaydigan sinflar orqali anglash masalasida tavsiflangan obyektlarning bog‘langanlik munosabatlaridan foydalanish taklif qilinadi.

T dagi chegaraviy ob'yektlar to'plami – $T_c \subset T$ bo'lsinki, yevklid metrikasi bo'yicha $T_c = \left\{ M_i \in Q_u | \rho(M, M_i) = \min_{M_j \in Q_u, M \in Q_{3-u}} \rho(M, M_j) \right\}$, $r_i = \min_{M_i \in Q_u, M \in Q_{3-u}} \rho(M, M_i)$, $r_j = \min_{M_j \in Q_u, M \in Q_{3-u}} \rho(M, M_j)$. M_i, $M_j \in Q_u$ ob'yektlar $M \in Q_u \cap T_c$ bo'yicha bog'langan hisoblanadi, agar $\rho(M_i, M) < r_i$ va $\rho(M_j, M) < r_j$. Bog'langanlik munosabati T to'plamning Ψ_1, ..., Ψ_α, $\alpha \leq h$ o'zaro kesishmaydigan guruhlarga bo'linishining yagonalini kafolatlaydi.

Mavzularga bo'lish ma'lum bir sinf hujjatlari uchun xarakterlanuvchi so'zlar to'plamini ajratish imkonini beradiki, ularning boshqa sinfda qo'llanilish ehtimolligi nisbatan past bo'ladi. Guruhlarning optimal soni quyidagicha

$$k_{opt} = \arg\max_t F(t, Y(guruh))$$

aniqlanadi.

K_1 sinf "fizika + matematika" fanlarining hujjatlari bilan ifodalangandagi mazmunning haqqoniyligini hisoblash natijalari 6-jadvalda ko'rsatilgan.

6-jadval.

Mazmunning haqqoniylik qiymatini hisoblash natijalari

Klasterlar soni	Mazmunning haqqoniyligi	Klasterlar soni	Mazmunning haqqoniyligi
2	0.7968	11	0.8898
3	0.8886	12	0.8794
4	0.8800	13	0.8892
5	0.8794	14	0.8861
6	0.8953	15	0.8916
7	0.8910	16	0.8916
8	0.8965	17	0.8910
9	0.8953	18	0.8904
10	0.8910		

Iyerarxik algoritmga alternativ variant sifatida (6) ketma-ketlikni keyinchalik ulardan latent alomatlarni shakllantirish bilan "ekspert" tomonidan berilgan sondagi guruhlarga ajratish taklif qilingan. Ushbu variantni tushuntirish uchun "kimyo" + "texnologiya" (K1) va K2 – qolgan 10 ta fan sohalarining hujjatlar to'plamidan foydalanish taklif etiladi.

Turg'unligi (13) bo'yicha tartiblangan 120 ta so'zdan iborat to'plam (umumlashtirilgan lug'at), ulardan latent alomatlarni shakllantirish maqsadida 6 ta guruhga bo'lingan. Y(6) to'plam bo'yicha K – means algoritmi yordamidagi klaster tahlilning natijalari 7-jadvalda keltirilgan.

7-jadval.

Klaster tahlilda mazmunning haqqoniyligi

Guruhlar soni	To'plamdagi alomatlar soni				
	2	3	4	5	6
2	0,8765	0,8680	0,8680	0,8680	0,8680
6	0,8765	0,8810	0,8810	0,8810	0,8810
19	0,8765	0,8810	0,8810	0,8810	0,8810
25	0,8765	**0,9026**	**0,9026**	**0,9026**	**0,9026**
37	0,8771	**0,9039**	**0,9039**	**0,9039**	**0,9039**

7-jadvaldan ko‘rinib turibdiki, turli xil sondagi guruhlar uchun mazmunning haqqoniyligi (14) qiymatlari juda yaqin. Ushbu yaqinlik dastlabki 3 ta latent alomat bo‘yicha kompaktlik o‘lchamining (3) qiymatlari yuqoriligi bilan izohlanadi.

8-jadval.

K_1 (Kimyo_Texnologiya) sinfdagi mavzular bo‘yicha so‘zlar ketma-ketligi

№	So‘zlar ketma-ketligi	№	So‘zlar ketma-ketligi
1	Nukleofil	9	Elektrod
2	Yuridik	10	Adaptasiya
3	Badiiy	11	Gibrid
4	Fosfor	12	digidroxinazolin
5	Potensial	13	Preparat
6	Matrisa	14	Reabilitasiya
7	Fragment	15	Magnit
8	Analitik	16	Gipertenziya

Hujjatlarning K_1 dagi mavzu bo‘yicha so‘zlar ketma-ketligi (8-jadvalga qarang) 12 ta fan bo‘yicha lug‘atlarni birlashtirish orqali olingan va ikkita o‘zaro kesishmaydigan sinflarga bo‘lingan ilmiy ishlarning o‘ziga xos xususiyatlarini ma’lum darajada aks ettiradi.

3.2-§da “kimyo + texnologiya” sinfi mavzulari bo‘yicha terminlarning umumiy lug‘atlari va bog‘langanlik munosabatlari bo‘yicha ketma-ketliklarini shakllantirishda o‘zaro kesishmaydigan 3 ta klaster olingan. Har bir klaster uchun atamalarning uchrash chastotalariga qarab tartiblanishi (tartiblangan) keltirilgan bo‘lib, natijalar 9-jadvalda ko‘rsatilgan.

T mavzulardan tematik model qurish masalasi nokorrekt qo‘yilgan hisoblanganligidan, uning yechimi uchun akademik Tixonov qoidalariga asosan regulyarizatsiyalashni qo‘llash lozim. 3.3-§da regulyarizatorlar matematik lingvistikada hal etilayotgan masalaning o‘ziga xos talablarini hisobga oladigan qo‘shimcha optimallik kriteriyalari sifatida qaraladi.

Tematik modelni qurish masalasi nokorrekt qo‘yilganligi sababli, uni hal qilish uchun akademik Tixonovning qoidalariga asoslanib, regulyarizatsiyani qo‘llash kerak bo‘ladi. Regulyarizatorlar yordamida mavzularni izohlashning ma’nosi quyidagicha.

1. $B \subset T$ fon mavzularini silliqlash.
2. Fan mavzularini siyraklash $S=T\backslash B$.
3. Mavzularning farqlanishini oshirish uchun dekorrelyatsiyalash.
4. Mavzularning izohlanishini yaxshilash uchun silliqlash + siyraklash + dekorrelyatsiyalash.

Qo‘shimcha ma’lumotlarni hisobga olish uchun regulyarizatorlar quyidagicha ko‘rinishda ifodalangan.

1. Vaqt modalligi bilan vaqtinchalik modellar.
2. Chiziqli regressiya modeli.
3. So‘zlarning mosligi modellari.
4. Ajdod mavzularni avlod qism mavzular bilan bog‘liqligi.

Amaliy masalalarda regulyarizatorlarni kombinatsiyalashga misollar:
– ijtimoiy tarmoqlardagi etnorelevant diskurslarni aniqlash;
– ilmiy va ilmiy-ommabop maqolalarni mavzu bo‘yicha izlash;
– yangiliklar oqimidagi hodisalarni aniqlash va kuzatish.

9-jadval.

Klasterlardagi terminlar rangi

Klaster		
1	2	3
ilmiy 5.1872	ilmiy 3.3079	kon 2.5673
disserta 3.578	disserta 2.4592	ilmiy 2.3278
kon 3.2724	kon 2.2433	texnologi 1.983
ta’lim 2.5115	texnologi 1.4579	disserta 1.8233
foyda 2.0863	foyda 1.2714	foyda 0.9276
davlat 1.9492	model 1.053	model 0.8049
huquq 1.9125	doktor 0.9065	energiya 0.7321
iqtisod 1.862	spektr 0.8578	texnika 0.7202
texnologi 1.4971	baho 0.7533	doktor 0.6756
baho 1.4962	institut 0.741	harorat 0.566
doktor 1.4407	harorat 0.6385	institut 0.5373
pedagog 1.3692	funksiya 0.6312	tezli 0.5249
o‘qit 1.021	tenglama 0.6297	konstruk 0.5164
qonun 1.0025	elektron 0.5968	dinamik 0.4661
institut 0.9021	davlat 0.578	koeffisiyent 0.4268
moliya 0.9017	foto 0.5705	mexanik 0.4235
mexanizm 0.7699	operator 0.5497	mashina 0.4136
referat 0.7503	energiya 0.5421	polimer 0.3935
model 0.7259	ta’lim 0.5044	tenglama 0.3891
milliy 0.6924	atom 0.5033	resurs 0.3875
invest 0.6905	referat 0.4997	referat 0.3651
kompleks 0.6213	o‘tkazgich 0.449	metall 0.3448
statistik 0.6021	kislota 0.4358	davlat 0.3289
foiz 0.5785	iqtisod 0.4121	massa 0.3184
resurs 0.5756	kompleks 0.4111	baho 0.3165
malaka 0.573	molekul 0.3875	oksid 0.3113
nazariy 0.5667	texnika 0.3822	kislota 0.3002
pul 0.5472	yadro 0.3822	formatsiya 0.2868
bozor 0.5389	ma’dan 0.3805	deformatsiya 0.2743
soliq 0.5327	resurs 0.3761	nazariy 0.2681
immun 0.5283	koeffitsiyent 0.3643	professor 0.2525
tarbiya 0.5028	birik 0.3569	iqtisod 0.2391
funksiya 0.4972	metall 0.3487	spektr 0.236
madani 0.4964	dinamik 0.3479	sinte 0.2318

Regulyarizator sifatida “o‘lcham la’natisi” sababli mashina algoritmlarini me’yoridan ortiq o‘rgatishdan qochish uchun qo‘llaniladigan terminlar fazosini siqish parametri taklif etiladi.

Regulyatsiya parametrlari, shuningdek, latent alomatlar soni va mavzularning mazmuniy haqqoniylik ko‘rsatkichi orqali ifodalanadi. Ishlab chiqilgan regulyarizatorlar ehtimollik talqiniga ega emas. Tadqiqot uchun latent

alomatlarning aniqlangan to‘plami bo‘yicha hujjatlar o‘rtasidagi munosabatlar qo‘llaniladi, keyinchalik ular xom alomatlarning kengaytirilgan (lekin cheklangan) fazosida qaraladi.

Xom alomatlar miqdoriga cheklov o‘zaro kesishmaydigan ikkita sinf hujjatlari uchun umumiy lug‘atning samarasi bilan ifodalanadi. Obyektlarning bog‘langanlik munosabatlari kalit tushuncha bo‘lib, u haqqoniylik mezoni bo‘yicha mavzularning optimal (lokal ma’noda) sonini tanlash masalasiga moslashadi.

Xulosa

“O‘zbek tilidagi hujjatlarning tematik klassifikatorlarini yaratish” mavzusidagi dissertatsiya tadqiqotining natijalari quyidagilardan iborat.

1. Turli fan sohalariga oid hujjatlarning semantik bog‘langanligini hisoblash formulasi taklif qilingan.

2. Tabiiy til terminlaridan latent fazoni shakllantirish texnologiyasi yaratilgan. Fazo hujjatlarni optimal sondagi mavzularga ajratish maqsadida klasterlashning metrik algoritmlarida qo‘llanilgan.

3. Hujjatlarning mazmuniy haqqoniylik bahosini hisoblash metodikasi ishlab chiqilgan. Metodika doirasida hujjatlar tavsiflarini haqqoniylik kriteriyasiga ko‘ra klasterlarga (mavzularga) optimal bo‘linishini tanlash taklif etilgan. Metodika mavzuga yo‘naltirilgan lug‘atlardagi so‘zlarni tanlashni asoslash, so‘zlarning omonimiyasi, polisemiyasini hisobga olgan holda bilimlar bazasini shakllantirish imkonini beradi.

4. Hujjatlarning ikkita kolleksiyasi uchun umumiy lug‘atlarni shakllantirish metodikasi ishlab chiqilgan. Metodika alomatlarni keyinchalik lug‘atga kiritish uchun ularning (terminlar chastotasi) turg‘unlik qiymatlarini hisoblash va tartiblashga asoslangan.

5. “Fizika + matematika” va “kimyo + texnologiya” fanlari bo‘yicha hujjatlar to‘plamidan semantik bog‘langan terminlar ketma-ketligi qo‘lga kiritilgan.

6. Hujjatlar sinflarining kompaktligini ularning o‘zaro bog‘langanlik munosabatlariga ko‘ra baholash texnologiyasi ishlab chiqilgan. Hujjatlarning o‘zaro bog‘langanlik munosabatlari hujjatlarning klaster tuzilishini tahlil qilish vositasi va tematik modellashtirishda yangi bilimlar manbai bo‘lib xizmat qilgan.

7. Hujjatlar sinflarining semantik bog‘langanlik darajasini hisoblash usuli aniqlangan. Hisoblash g‘oyasi ikkita sinfdagi hujjatlarni bittaga birlashtirganda, ularning kompaktlik o‘lchami har birini alohida qaralganidan kattaroq bo‘lishi haqidagi gipotezani tekshirishga asoslanadi.

8. Turli predmet sohalariga mansub so‘zlarning omonimiya, ko‘p ma’nolilik muammosini latent alomatlar fazosini shakllantirish orqali hal etish taklif etilgan. Latent alomatlarga o‘tish tahlil qilish uchun fazoning hajmini kamaytirish, informativligi kam alomatlarni olib tashlash va predmet sohalar uchun lug‘atlarni shakllantirish jarayonini avtomatlashtirish imkonini beradi.

9. Keyingi tadqiqotlar istiqbollari semantik bog‘langanlikni baholash texnologiyalaridan foydalanish va tematik hujjatlarning barcha uslublari uchun atamalarning umumiy lug‘atlarini yaratish bilan bog‘liq.

НАУЧНЫЙ СОВЕТ DSc.03/30.12.2019.FM.01.02
ПО ПРИСУЖДЕНИЮ УЧЕНЫХ СТЕПЕНЕЙ ПРИ
НАЦИОНАЛЬНОМ УНИВЕРСИТЕТЕ УЗБЕКИСТАНА

НАЦИОНАЛЬНЫЙ УНИВЕРСИТЕТ УЗБЕКИСТАНА

ТУЛИЕВ УЛУГБЕК ЮЛДАШЕВИЧ

РАЗРАБОТКА ТЕМАТИЧЕСКИХ КЛАССИФИКАТОРОВ ДОКУМЕНТОВ НА УЗБЕКСКОМ ЯЗЫКЕ

05.01.11 – Цифровые технологии и искусственный интеллект

АВТОРЕФЕРАТ
диссертации доктора философии (PhD) по
ФИЗИКО-МАТЕМАТИЧЕСКИМ НАУКАМ

Ташкент – 2023

Тема диссертации доктора философии (Doctor of Philosophy) по физико-математическим наукам зарегистрирована в Высшей аттестационной комиссии при Министерстве высшего образования, науки и инноваций Республики Узбекистан за № B2022.4.PhD/FM831.

Диссертация выполнена в Национальном Университете Узбекистана имени Мирза Улугбека.

Автореферат диссертации на трех языках (узбекский, русский, английский (резюме)) размещен на веб-странице Научного совета (http://ik-fizmat.nuu.uz/) и на Информационно-образовательном портале «Ziyonet» (www.ziyonet.uz).

Научный руководитель:	**Игнатьев Николай Александрович** доктор физико-математических наук, профессор
Официальные оппоненты:	**Матякубов Алишер Самандарович** доктор физико-математических наук, доцент
	Норов Абдусаид Муродович PhD по технических наук
Ведущая организация:	**Ургенчский государственный университет**

Защита диссертации состоится «____» _____2023 года в ____ часов на заседании Научного совета DSc. 03/30.12.2019.FM.01.02 при Национальном университете Узбекистана. (Адрес: 100174, г. Ташкент, Алмазарский район, ул. Университетская, 4. Тел.: (+99871)227-12-24, факс: (+99871) 246-53-21, e-mail: nauka@nuu.uz).

С диссертацией можно ознакомиться в Информационно-ресурсном центре Национального университета Узбекистана (зарегистрирована за №_____). (Адрес: 100174, г. Ташкент, Алмазарский район, ул. Университетская, 4. Тел.: (+99871) 246-02-24).

Автореферат диссертации разослан «____» _____2023 года.
(протокол рассылки №_____ от «____» _____2023 года).

М.М.Арипов
Председатель Научного совета по присуждению ученых степеней, д.ф.-м.н., профессор

З.Р.Рахмонов
Ученый секретарь Научного совета по присуждению ученых степеней, д.ф.-м.н.

Д.Т.Мухамедиева
Председатель научного семинара при Научном совете по присуждению ученых степеней, д.т.н., профессор

ВВЕДЕНИЕ (аннотация диссертации доктора философии(PhD))

Актуальность и востребованность темы диссертации. Многие научно-прикладные исследования ведущиеся в мире, в настоящее время связаны с такими вопросами, как разработка методов анализа текста (text Mining) и обработки естественного языка. Разработка методов анализа текстов (text mining) получило широкое распространение в мире. Язык является самым мощным средством для передачи знаний между людьми. Серьезным препятствием для передачи знании является в различие в синтаксисе и семантике естественных языков. Для формирования и усвоения синтаксиса и семантики используются разные методы. Смысл языка пытаются изучать путем создания гигантских баз данных (корпусов) для хранение документов на естественном языке. Процесс извлечения знаний из описаний документов представляет решение ряда задач в рамках бурно развивающихся областей как интеллектуальный анализ данных (data mining). Поэтому представление текстовых документов на естественном языке в виде числовых векторов, разработка тематических моделей на основе их семантических связей остается одной из важных задач интеллектуального анализа данных.

В настоящее время в мире собрано огромное количество научных работ по машинной лингвистике, которые позволяют на практике использовать методы и алгоритмы работы с семантикой естественного языка. Особенности естественного языка приходится учитывать при разработке их моделей. Получены оригинальные результаты в области методов, алгоритмов и инструментальных средств для решения проблем Big Data, представленных в естественном языке. Именно большие данные стали катализатором появление новых научных знаний и технологических решений, связанных с развитием семантических технологий. В частности методы обработки естественного языка направлены на разрешения многозначности в различных его проявлениях.

Появления новых методов ИАД делает востребованным решение задач семантической связанности, вычисления контентной аутентичности документов, автоматизации построения словарей для предметных областей. В частности, если в задачах приобретения знаний подразумевается, что представление знаний является заданными априори и нужно лишь построить систему в рамках этих представлений, то в задачах метаобучения ставится вопрос об автоматическом представлении самих представлений, детали которых могут сильно меняться в зависимости от предметной области. Решение задач метаобучения необходимо для снятия следующего ограничения для машинных систем – их способности функционировать только в узкой предметной области.

В нашей стране в годы независимости большое внимание уделяется научным направлениям, имеющим фундаментальное и прикладное значение. В качестве приоритетов определены формирование большого объема цифровых данных на государственном языке для машинного обучения

искусственного интеллекта, а также разработку программных продуктов, использующих анализ и синтез речи на государственном языке[1]. Эти задачи потребовали решения таких задач, как автоматизация построения словарей по предметным областям, расчет достоверности содержания документов, семантической связи. В математической лингвистике актуальной является разработка и использование онтологий предметных областей в виде словарей, тезаурусов и глоссариев на узбекском языке, для создания которых используются методы машинного обучения. На основе Стратегии Действий по развитию Республики Узбекистан особенно большое значение приобретают эффективные механизмы внедрения научных и инновационных достижений в целях повышения эффективности в сфере экономики страны.

Этим целям служит в Указах Президента Республики Узбекистан №УП-4947 от 7 февраля 2017 г. «О Стратегии действий по дальнейшему развитию Республики Узбекистан», Постановлениях Президента Республики Узбекистан №ПП-2789 от 17 февраля 2017 г. «О мерах по дальнейшему совершенствованию деятельности Академии наук, организации, управления и финансирования научно-исследовательской деятельности», №ПП-2909 от 20 апреля 2017 г. «О мерах по дальнейшему развитию системы высшего образования», №ПП-3682 от 27 апреля 2018 г. «О мерах по дальнейшему совершенствованию системы практического внедрения инновационных идей, технологий и проектов», доклад Президента Республики Узбекистан Ш.Мирзиёева 24 мая 2019 года на встрече с представителями науки и образования в Национальном университете Узбекистана, его указ №-УП-6084 от 20 октября 2020 года «О мерах по дальнейшему развитию узбекского языка и совершенствованию языковой политики в стране», его постановление №-ПП-4996 от 17 февраля 2021 года «О мерах по созданию условий для ускоренного внедрения технологий искусственного интеллекта», а также в других нормативно-правовых актах по данной деятельности.

Соответствие исследования приоритетным направлениям развития науки и технологий республики. Данное исследование выполнено в соответствии с приоритетным направлением развития науки и технологий Республики Узбекистан IV. «Математика, механика и информатика».

Степень изученности проблемы. Востребованность решения задач математической лингвистики заметно ощутимо по наличию широко используемого на практике программного обеспечения для перевода с одного языка на другой. Например, системой Google Translate пользуются миллионы людей. К сожалению, результаты качества перевода еще очень далеки от того, что могут сделать профессионалы (человек). Предстоит большая работа по совершенствованию семантики перевода, созданию тематических корпусов документов на естественных языках. Другой проблемой является автоматизация процесса работы с тематическими документами. В системе государственного управления выдвигается и реализуются амбициозные

[1] Постановление Президента Республики Узбекистан, №-ПП-4996 от 17 февраля 2021 года «О мерах по созданию условий для ускоренного внедрения технологий искусственного интеллекта».

программы создания электронного правительства. Важный составляющей этих программ является создание унифицированных документов, однозначная интерпретация терминов и понятий для устранения разночтений в них.

Для относительно небольшого количества естественных языков разработаны системы проверки правописания. Использование таких систем на практике способствует оперативному обнаружению и исправлению ошибок. Существует подход к решению некорректных задач, называемый регуляризацией. Когда оптимизационная задача недоопределена, к основному критерию добавляют дополнительный критерий – регуляризатор, учитывающий специфику решаемой задачи и знания предметной области. В практических задачах автоматической обработки текстов происходит разработка дополнительных критериев и ограничений на решения, которых может быть много по причине их слабой формализации.

В развитие теории и практики использования методов распознавания большой вклад внесли известные зарубежные и отечественные учёные. Среди зарубежных учёных можно выделить Ю.И. Журавлева, Н.Г. Загоруйко, В.А. Дюк, К.В. Воронцова, М. Ливенштейна, среди отечественных особо следует отметить труды М.М. Камилова, Т.Ф. Бекмуратова, Ф.Т. Адыловой, Г.Р. Матлатипова, М. Мусаева и др.

Связь темы диссертации с научно-исследовательскими работами учреждением высшего образования, где выполнялась диссертация. Диссертационное исследование выполнено в рамках научного гранта согласно плану научно-исследовательских работ Национального университета Узбекистана Ф-4-64 «Разработка и обоснование методов вычисления обобщённых оценок и индивидуальных метрик объектов в интеллектуальном анализе данных» (2011-2016 гг.).

Целью исследования является разработка критериев и методов интеллектуального анализа данных для решения задач тематического моделирования.

Задачи исследования состоят в следующем:

автоматизировать процесс создания словарей для коллекций документов на естественном языке;

оценить компактность классов документов по отношению связанности;

определить степень семантической связанности документов в коллекции;

выделить искусственные паттерны из слов по парам классов документов;

разработать метод вычисления параметров контентной аутентичности документов.

Объект исследования. Разработка и обоснование технологий анализа семантики документов на естественных языках.

Предмет исследования. Методы метрической классификации для анализа коллекций документов в задачах тематического моделирования.

Методы исследования. Поиск оптимальных алгоритмов классификации на основе методов дискретной математики, теории искусственного интеллекта, математического анализа, программирования на алгоритмических языках.

Научная новизна исследования заключается в следующем:

предложено и обоснованно вычисление коэффициента семантической связанности документов на основе количественной оценки отношений описаний документов. Самое высокое значение коэффициента в документах с научным стилем изложения получено по предметам математика и физика;

выделены искусственные паттерны из слов по парам классов документов с использованием латентных признаков, сформированных методом вычисления обобщённых оценок по непересекающимся группам из сырых признаков;

разработана методика вычисления показателя контентной аутентичности коллекций документов. В методике использовались критерии для определения оптимального числа тем по результатам кластерного анализа;

разработан метод формирования общего словаря терминов для данных из двух коллекций документов. При отборе терминов в словарь используются показатели их устойчивости, вычисляемые по значениям функции принадлежности к классам.

Практические результаты исследования заключаются в возможности автоматической классификации документов по темам с учётом их семантической связанности.

Достоверность результатов исследования. Достоверность полученных результатов обосновывается тем, что итоги вычислительных экспериментов при решении задач тематического моделирования с использованием методов семантических вычислений подтверждают адекватность тематической классификации для предметно-ориентированных областей.

Научная и практическая значимость результатов исследования. Научная значимость результатов исследования заключается в разработке и обосновании методов вычисления коэффициентов связанности и семантической связанности классов и групп документов, формирования общих словарей для коллекций документов, алгоритме метода вычисления контентной аутентичности документов.

Практическая значимость заключается в разработке программного обеспечения для реализации алгоритмов методов семантического анализа, позволяющие получать новые знания и решать проблему многозначности смысла слов из разных предметных областей.

Внедрение результатов исследования. На основе научных результатов, полученных при разработке метода формирования общего словаря терминов для данных в коллекции документов путем расчета коэффициента их смысловой связи и показателя достоверности содержания коллекции документов на основе количественной оценки их взаимоотношений при описании документов:

результаты, полученные в диссертационной работе по созданию метода расчета семантической связи между наборами документов, были использованы при хронологическом анализе текстовых данных в грантовом

проекте «Создание веб-портала «Бахтиёр Оила» в Узбекистане» с номерам JHBL-13. (справка № 10 от 25 июня 2021 года НИИ «Махалля и семья»). Применение научных результатов позволило сформировать словари, организующие признакового пространство для классификации текстовых документов;

Результаты, полученные по формированию общего словаря коллекций документов и количественной оценке отношений описаний документов, были использованы при создании глоссариев терминов на узбекском языке и в системах электронного документооборота в Научно-исследовательском центре науки, технологий и маркетинг ГУП «UNICON.UZ» (справка № 1-1/300 от 20 февраля 2023 года ГУП «UNICON.UZ»). Программное обеспечение для разработки тематических моделей позволило повысить эффективность и оперативность принимаемых решений на 40% за счет автоматизации процесса создания словарей научных терминов и электронных документов на их основе.

Апробация результатов исследования. Результаты данного исследования были обсуждены на 6 научно-практических конференциях, в том числе на 4 международных и 2 республиканских.

Публикация результатов исследования. По теме диссертации опубликовано 16 научных работ, из них 6 входят в перечень научных изданий, предложенных Высшей аттестационной комиссией Республики Узбекистан для защиты диссертаций доктора философии, в том числе из них 3 опубликованы в зарубежных журналах и 3 в республиканских научных изданиях. А также получены 3 свидетельства регистрации программных продуктов, созданных для ЭВМ.

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

ОСНОВНОЕ СОДЕРЖАНИЕ ДИССЕРТАЦИИ

Во введении обоснована актуальность и востребованность темы диссертации, определено соответствие исследования приоритетным направлениям развития науки и технологий республики, приведены обзор зарубежных научных исследований по теме диссертации и степень изученности проблемы, сформулированы цели и задачи, выявлены объект и предмет исследования, изложены научная новизна и практические результаты исследования, раскрыта теоретическая и практическая значимость полученных результатов, даны сведения о внедрении результатов исследования, об опубликованных работах и о структуре диссертации.

В главе I **Модели данных естественного языка** рассматриваются особенности тематического моделирования. С этой целью в §1.1 приводится терминология для семантического анализа текстов на естественном языке. Поскольку семантика слов, понятий, паттернов ЕЯ определяется контекстом, то требуется объяснение их смысла с помощью других слов. Для этого используются различные виды онтологий. Эти виды содержат целевые разделения онтологий на:

- словарь – список однозначных терминов;

- глоссарий – словарь многозначных терминов с перечислением их значений;

- тезаурус – глоссарий с заданный системой семантических связей.

Описаны термины, наиболее часто применяемые в тематическом моделировании. Акцент сделан на понятия, используемые в онтологии для предметно-ориентированных словарей. Функциональная схема формирования словарей, предлагаемая в данной работе показана на рис. 1.

Рис. 1. Процесс формирования словарей

В §1.2 рассматривается кластерный анализ документов с использованием алгоритмов классификации. Кластеризация множества текстовых документов, представляет разбиение их на близкие по смыслу подмножества.

При кластеризации признаков во внимание принимается проблема BigData, известная как проклятие размерности. Процесс кластеризации реализуется через вычисления латентных признаков по непересекающимся группам сырых методом обобщённых оценок объектов. Схема исследование показано в рис.1.

Перечень традиционных методов был дополнен алгоритмами группировки, использующими отношение попарной связанности документов на основе логической закономерности в форме гипершаров.

Целью формирования баз тематических документов в § 1.3 является анализ закономерностей и особенностей ЕЯ. Набор текстов и дополнительная информация об этих текстах в коллекциях документов, написанных на одном языке, образует корпус.

Корпус важен для анализа и создания модели ЕЯ. Корпус содержит документы, написанные разным стилем речи (публицистический, научный, официальный, художественный). Поскольку понятий больше, чем слов, то слова могут использоваться в разных предметных областях со своим

смыслом. В этом случае семантические отношения между словами могут помочь выбрать правильный смысл. Кроме того, слово может дать больше смысла, если оно идет в связи с другим словом. Этим целям служит предметно-ориентированный словарь, как набор лексики, организованной по тематическому (семантическому) принципу с отражением совокупности базовых семантических отношений определенной предметной области.

В главе II **Тематическое моделирование по коллекциям документов с научным стилем написания** рассматриваются особенности анализа документов с научным стилем написания. В §2.1 научный стиль представлен через анализ авторефератов научных диссертаций из коллекции ВАК Республики Узбекистан.

Узбекский язык является примером малораспространенного языка с точки зрения представления результатов научных исследований, поскольку открытых источников данных для машинной обработки в виде коллекций текстовых документов сравнительно мало. Одним из требований к этим документам является то, что они должны быть написаны с соблюдением правил узбекской грамматики. Смысл предобработки текстов во многом сводится к приведению всех слов к нормальной форме через лемматизацию или стемминг. Побудительными мотивами для работы с коллекциями авторефератов были следующие.

1. Не существует официально утвержденный и доступный для массового использования корпус узбекского языка. Открытым остаётся вопрос выбора нормальной формы предобработки слов для последующего их использования в машинном обучении.

2. Содержимое авторефератов приводится на трех языках узбекском, русском или английском со сходной структурой изложения материала диссертации. Наличие корпусов русского и английского языков и методов анализа текстов на их основе позволяет проводить сравнение содержимого авторефератов на узбекском языке.

3. Поиск связей между различными научными дисциплинами для объяснения причин изменения количества тем и их содержимого при тематическом моделировании.

4. Формирование базы экспертов, рецензентов для оценки содержимого диссертаций.

Для экспериментов использовались 1634 документа, разбиение которых по предметным областям приводится в табл. 1.

Таблица 1.

Разбиение документов по предметным областям

Предметная область	Количество документов	Предметная область	Количество документов
Биология	109	Юриспруденция	83
Физика	162	Экономика	189
География	28	Химия	120
Геология	44	Культура	11
Математика	95	Техника	380
Педагогика	137	Медицина	277

Вычисление меры компактности и меры связанности тематических документов в §2.2 реализуется исходя из следующей постановки задачи.

Считается, что задано множество объектов $E_0=\{S_1,\ldots,S_m\}$, разделённое на l ($l\geq 2$) непересекающихся подмножеств (классов) $K_1,\ldots,K_l$. Описание объектов производится с помощью набора из n количественных признаков $X(n)=(x_1,\ldots,x_n)$**.** На множестве описаний объектов E_0 задана метрика $\rho(x,y)$ и определена мера компактности классов $\mu(E_0,\rho,R^n)$. Требуется:

-определить набор $X(d)\subset X(n)$, $d\leq n$ без шумящих признаков;

-реализовать отображение описаний документов из пространства R^d в пространство размерности R^k, $R^d\rightarrow R^k$, $k<d$;

-определить оптимальные размеры пространства R^d по

$$\mu(E_0,\rho,R^n)\,\mathrm{d}=\arg\max(t\in\{2,\ldots,n\})\,\mu(E_0,\rho,R^t)$$

Значения меры компактности (см. табл.2) класса K_t, $t=1,\ldots,l$ вычисляются через разбиение объектов по отношению связанности по системе гипершаров на непересекающихся группы $G_{t1},\ldots,G_{t\beta(t)}$, $\beta(t)\geq 1$ по

$$\Theta_t=\frac{\sum_{i=1}^{\beta(t)}\left|G_{ti}\right|^2}{\left|K_t\right|^2}, \qquad (1)$$

и выборки в целом по

$$\Theta=\frac{\sum_{i=1}^{l}\left|K_i\right|\Theta_i}{m}. \qquad (2)$$

Таблица 2.

Значения меры компактности документов по предметным областям

Предметная область	Мера компактности (1)	Предметная область	Мера компактности (1)
Биология	0.4370	Химия	0.3655
Физика	0.0515	Культурология	0.6859
География	0.2976	Математика	0.0145
Геология	0.5103	Педагогика	0.9426
Юриспруденция	0.9294	Техника	0.5138
Экономика	0.7829	Медицина	0.8678

Компактность (2) по выборке равна 0.5686. Для оценки сходства терминологии из разных предметных областей в §2.3 поставлена следующая задача.

Считается, что задано множество объектов $E_0=\{S_1,\ldots,S_m\}$, разделённое на l ($l>2$) непересекающихся подмножеств (классов) $K_1,\ldots,K_l$, $E_0=\bigcup_{i=1}^{l}K_i$. Описание объектов (документов) производится с помощью набора из n количественных признаков $X(n)=(x_1,\ldots,x_n)$. На множестве описаний объектов задана метрика $\rho(x,y)$, определена мера компактности $\Theta(K_i,E_0,\rho)$, $\Theta(K_j,E_0,\rho)$ и объединения классов $K_i\cup K_j$ $\Theta(K_i\cup K_j,E_0,\rho)$, $i,j\in\{1,\ldots,l\}$. Требуется определить

попарную меру сходства и меру связаности классов по $\Theta(K_i,E_0,\rho)$, $\Theta(K_j,E_0,\rho)$, $\Theta(K_i \cup K_j,E_0,\rho)$.

Пусть $\theta_i, \theta_j, \theta_{ij}$- значения меры компактности документов соответственно из K_i, K_j и $K_i \cup K_j$, $i,j \in \{1,\ldots,l\}$ в $\cup_{d=1}^{l} K_d$, $l>2$, определяемые по (1), (2). Тогда попарная мера сходства или семантическая близость документов $\varphi(K_i, K_j)$ в $\cup_{d=1}^{l} K_d$ будет вычисляться как

$$\Phi(K_i, K_j) = 1 - \frac{\theta_i}{\theta_{ij}} \frac{\theta_j}{\theta_{ij}}. \quad (3)$$

Выделим по (3) множество пар классов объектов

$$\Omega=\{(K_i,K_j)|\varphi(K_i,K_j)>0,\ i,j\in\{1,\ldots,l\}\} \quad (4)$$

и определим отношение

$$R(K_i, K_j) = \frac{\max(\theta_i,\theta_j)}{\theta_{ij}} \quad (5)$$

как семантическую связанность документов.

Таблица 3.

Попарная мера сходства по научным дисциплинам

Документы по дисциплинам	Мера компактности		$1 - \frac{\theta_i}{\theta_{ij}} \frac{\theta_j}{\theta_{ij}}$
	θ_i, θ_j	θ_{ij}	
Математика, физика	0.0145	0.3357	0.9896
	0.0804		
Математика, медицина	0.0145	0.5237	0.9537
	0.8745		
Математика, техника	0.0145	0.5253	0.9440
	0.5176		
Математика, экономика	0.0145	0.3484	0.9063
	0.7829		
Математика, педагогика	0.0145	0.3361	0.8788
	0.9426		
…			
География, физика	0.2857	0.0647	-4.4911
	0.0804		
Геология, физика	0.5103	0.0749	-6.3228
	0.0804		
География, математика	0.2857	0.0235	-6.5322
	0.0145		
Культурология, физика	0.6859	0.0733	-9.2629
	0.0804		
Культурология, математика	0.6859	0.019	-26.4481
	0.0145		

При объединении документов по (см. табл. 3) в один класс была получено значение меры компактности, равное 0.3357. Данный результат показывает наличие родственных терминов по этим предметам, которые

нашли отражение при выборе научной классификации для присуждения ученых степеней по физико-математическим наукам.

Сформированная на основе (3), (4) и упорядоченная по (5) последовательность множеств документов по парам дисциплин приводится в табл. 4.

Таблица 4.

Семантическая связанность документов

Дисциплины	Отношение (5)	Дисциплины	Отношение (5)
Математика, физика	0.2397	Биология, медицина	0.9672
Химия, техника	0.7259	Юрисприденция, педагогика	0.9869
География, геология	0.8967	Биология, геология	0.9933
География, техника	0.9122	Химия, биология	0.9945
Геология, техника	0.9298	География, экономика	1.0084

В главе III **Контентная аутентичность документов** рассматривается вычисление показателей контентной аутентичности документов для кластерного анализа данных, проводимого с использованием метрических алгоритмов. Как одной из задач, связанной с Big Data, приводится отбор термов в общие словари документов.

В §3.1 разработка искусственных паттернов в задачах математической лингвистики проводится из термов ЕЯ. Паттерны формируются как латентные признаки или метапризнаки относительно пары коллекций документов.

Постановка задачи. Дана выборка объектов $E_0=\{S_1,\ldots,S_m\}$, разделённая на 2 непересекающихся класса K_1, K_2. Описание объектов в E_0 представлены набором количественных признаков $X(n)=(x_1,\ldots,x_n)$. Определены процедуры для:

– разбиения $X(n)$ на непересекающиеся подмножества $X(n_1),\ldots,X(n_t)$ по классам K_1,K_2, $t\geq 1$, $\forall c\in\{1,\ldots,t\}$ $n_c>1$, $n_1+\ldots+n_t\leq n$ и вычисления по ним набора латентных признаков $Y(t)=(y_1,\ldots,y_t)$;

– разбиения объектов E_0 на заданное число непересекающихся кластеров (групп) $G_1,\ldots,G_p$;

– вычисления оценки качества разбиения $F(p)$ на заданное число p групп.

Требуется:

– выделить набор латентных признаков $Y(\sigma)\subset Y(t)$ для разбиения E_0 на p групп;

– определить число групп $\mu=\arg\max_p F(p)$;

–по каждой из групп $G_1,\ldots,G_\mu$ вычислить ранги слов из множества

$\Omega=\left\{X(n_c)\middle| y_{n_c}\in Y(\sigma)\right\}$;

– определить ранги наиболее употребительных слов из представителей двух классов документов.

Пусть для значений признака $x_c \in X(n)$ в описании объектов $K_1 \cup K_2$ построена упорядоченная по неубыванию последовательность

$$r_1,\ldots,r_j,\ldots,r_m,\ m=|K_1 \cup K_2|. \quad (6)$$

Разбиение (6) производится по двум интервалам $[c_1,\ c_2]$ и $(c_2,\ c_3]$, границы которых определены по критерию

$$\left(\frac{\sum_{p=1}^{2}\sum_{i=1}^{2} u_i^p\left(m-|K_i|-\sum_{j=1}^{2} u_j^p-u_i^p\right)}{2|K_1||K_2|}\right)\left(\frac{\sum_{p=1}^{2}\sum_{i=1}^{2} u_i^p\left(u_i^p-1\right)}{|K_1|(|K_1|-1)+|K_2|(|K_2|-1)}\right) \to \max_{c_1<c_2<c_3}, \quad (7)$$

где u_i^p – число объектов класса K_i в p-ом интервале.

Значения в границах интервала $[r_u;r_v]^i$, определяемые на (6) с помощью критерия

$$\left|\frac{d_{t,c}(u,v)}{|K_t|}-\frac{d_{3-t,c}(u,v)}{|K_{3-t}|}\right| \to \max \quad (8)$$

рассматриваются как градация номинального признака.

Пусть $d_{tc}(u,v)$, $d_{3\text{-}t,c}(u,v)$ – количество представителей классов K_t, $K_{3\text{-}t}$ в интервале $[r_u;r_v]^i$, $i \in \{1,\ldots,p_c\}$. Значение функции принадлежности $f_c(\mu)$ к классу K_1 по интервалу $[r_u;r_v]^\mu$(градации $\mu \in \{1,\ldots,p_c\}$) вычисляется как

$$F_c(\mu)=\frac{d_{1c}(\mu)/|K_1|}{d_{1c}(\mu)/|K_1|+d_{2c}(\mu)/|K_2|}. \quad (9)$$

Для вычисления обобщённых оценок объектов на E_0 используются вклады градаций признаков. Вклад градации $j \in \{1,2\}$ признака $x_c \in X(n)$ определяется как

$$\eta_c(j)=w_c\left(\frac{\alpha_{cj}^1}{|K_1|}-\frac{\alpha_{cj}^2}{|K_2|}\right), \quad (10)$$

где α_{cj}^1, α_{cj}^2 – количество значений градации j признака x_c соответственно в классах K_1 и K_2, w_c – вес признака x_c, определяемый как произведение внутриклассового сходства и межклассового различия. Обобщённая оценка объекта $S_r \in E_0$ по описанию в номинальной шкале измерений $S_r=\{a_{ri}\}_{i \in X(n)}$ на наборе $X(n)$ и вкладам (10) вычисляется как

$$R(S_r)=\sum_{i \in \{1,\ldots,n\}} \eta_i(a_{ri}). \quad (11)$$

Пусть в описании объекта $S_r \in K_1 \cup K_2$ исходные значения признаков из $X(n)$ заменены на значения функции принадлежности $S_r = \{b_{ri}\}_{i\in\{1,\dots,n\}}$ по (9). Устойчивость признака $x_c \in X(n)$ вычисляется как

$$\varphi(c) = \frac{1}{m}\sum_{r=1}^{m} \begin{cases} b_{rc}, b_{rc} > 0.5, \\ 1 - b_{rc}, b_{rc} < 0.5. \end{cases} \quad (12)$$

Упорядоченность слов по (12) в общем–то не зависит от размерности пространства для описания коллекции документов.

Смысл вычисления коэффициента контентной аутентичности заключается в поиске оптимального числа кластеров, определяющих разбиение документов на темы в латентном признаковом пространстве. Для формирования латентного пространства и общего словаря на его основе предлагается:

– упорядочить признаки с индексами из D по неубыванию их значений устойчивости (12) как

$$\varphi(\varepsilon_1), \dots, \varphi(\varepsilon_j), \dots, \varphi(\varepsilon_{dim}), \varepsilon_j \in D, dim = |D|; \quad (13)$$

– сформировать множество латентных признаков $Y(t)=(y_1,\dots,y_t)$ по (13);

– выделить $Y(\sigma) \subset Y(t)$, $\sigma \le t$ и множество, связанных по D с $Y(\sigma)$ сырых признаков в качестве общего словаря.

Количество слов для общего словаря из (13) является в общем–то свободным параметром. Этот параметр может задаваться по эвристическим соображениям, либо по результатам иерархической агломеративной группировки при формировании набора латентных признаков.

Обозначим через *TYPLAM* – множество индексов сырых признаков, включённых в состав группы алгоритмом иерархической группировки, *lugat* – ограничение на количество слов в общем словаре, *guruh* – количество латентных признаков. Реализация алгоритма по шагам будет следующей.

Шаг 1. j=0. *guruh*=0.

Шаг 2. Вычислить $j = j + 1$. *crit*=10. $u = \varepsilon_j$. *TYPLAM* = {u}. *guruh*=*guruh*+1.

Цикл по $t \in \{1,\dots,m\}$ $R(S_t) = \eta_u(a_{tu})$. Конец цикла;

Шаг 3. $u = \varepsilon_{j+1}$. Цикл по $t \in \{1, \dots, m\}$ $b_t = R(S_t) + \eta_u(a_{tu})$. Конец цикла;

$M_1 = \sum_{S_t \in K_1} b_t$. $M_2 = \sum_{S_t \in K_2} b_t$. $M_1 = M_1/|K_1|$. $M_2 = M_2/|K_2|$. θ=0. γ=0. Цикл по $t \in \{1, \dots, m\}$ Если $S_t \in K_1$, то $\theta = \theta + |b_t - M_1|$, $\gamma = \gamma + |b_t - M_2|$. Иначе $\theta = \theta + |b_t - M_2|$, $\gamma = \gamma + |b_t - M_1|$. Конец цикла;

Шаг 4. Если $\theta/\gamma < crit$, то $crit = \theta/\gamma$, *TYPLAM* = *TYPLAM* $\cup \{u\}$, $j = j + 1$, идти 3.

Шаг 5. Вывод $\{R(S_t)\}_{t\in\{1,\dots,m\}}$, *TYPLAM*.

Шаг 6. Если $j < lugat$, то идти 2; Иначе вывод *guruh*.

Шаг 7. Конец.

Демонстрация формирования латентных признаков алгоритмов приводится в табл.5. Класс K_1 представлен документами по физике и математике, K_2 – всеми остальными документами из 10 предметных областей.

Таблица 5.

Латентные признаки и их состав из сырых признаков.

№	Кол-во признаков	Признаки	Значения критерия (7)
1	267	179, 55, 171, … , 414, 297, 299	0.6360
2	87	423, 66, 450, … , 186, 36, 265	0.5061
3	66	70, 67, 192, …. , 290, 15, 21	0.3917
4	10	393, 61, 75, 253, 163, 405, 53, 24, 44, 490	0.2793
5	23	322, 408, 295, … , 311, 364, 316	0.3493
6	1	89	0.2442
7	16	139, 189, 166, … , 305, 199, 304	0.3492
8	8	50, 158, 223, 195, 427, 302, 433, 369	0.2372
9	2	98, 197	0.2127
10	1	177	0.1852
11	2	57, 279	0.2344
12	2	120, 135	0.2460

Пусть $G_1,...,G_h$, $i=1, ...,h$, $h\geq 2$ разбиение на непересекающиеся группы по набору латентных признаков $Y(guruh)$. Для каждой группе G_i определим значение функции принадлежности объектов к классу K_1 по G_i как $\lambda_i(K_1)=d_{i1}/|G_i|$, где d_{i1}– число объектов класса K_1 в G_i. Контентная аутентичность документов из E_0 при разбиении их на h групп будет вычисляться как

$$F\left(h,Y(guruh)\right)=\frac{1}{m}\sum_{j=1}^{h}\begin{cases}|G_j|\lambda_j(K_1),\lambda_j(K_1)>0.5;\\ |G_j|(1-\lambda_j(K_1)),\lambda_j(K_1)<0.5.\end{cases} \quad (14)$$

Целесообразность использования (14) для вычисления контентной аутентичности по общему словарю можно проверить следующим образом.

Будем считать, что известно разбиение объектов на группы $G_1, ...,G_h$ и представители какого класса K_1 или K_2 доминируют в каждой из них. Обозначим через $T=(M_1, ...,M_h)$ – множество центров групп (объектов), представленное как объединение двух непересекающихся подмножеств Q_1 и Q_2, $T=Q_1\cup Q_2$ с доминированием в группах представителей соответственно класса K_1 и K_2. Для определения оптимального числа подмножеств центров групп (тем документов) предлагается использовать отношение связанности объектов по классам.

Пусть $T_c\subset T$ – множество граничных объектов на T, $T_c=\left\{M_i\in Q_u|\rho(M,M_i)=\min_{M_j\in Q_u,M\in Q_{3-u}}\rho\left(M,M_j\right)\right\}$, $r_i=\min_{M_i\in Q_u,M\in Q_{3-u}}\rho(M,M_i)$, $r_j=\min_{M_j\in Q_u,M\in Q_{3-u}}\rho\left(M,M_j\right)$ по евклидовой метрике. Объекты M_i, $M_j\in Q_u$ связаны по $M\in Q_u\cap T_c$, если $\rho(M_i,M\ )<r_i$ и $\rho(M_j,M\ )<r_j$. Отношение связанности гарантирует единственность разбиения множества T на непересекающиеся группы $\Psi_1, ..., \Psi_\alpha$, $\alpha\leq h$.

Оптимальное число групп определяется как

$$k_{opt} = \arg\max_{t} F\left(t, Y(guruh)\right).$$

В табл. 6 представлена результаты вычислений контентной аутентичности, где K_1 представлен документам из предметов «физика + математика».

Таблица 6.

Результаты вычисления значений контентной аутентичности

Количество кластеров	Контентная аутентичность	Количество кластеров	Контентная аутентичность
2	0.7968	11	0.8898
3	0.8886	12	0.8794
4	0.8800	13	0.8892
5	0.8794	14	0.8861
6	0.8953	15	0.8916
7	0.8910	16	0.8916
8	0.8965	17	0.8910
9	0.8953	18	0.8904
10	0.8910		

В качестве альтернативного варианта иерархическому алгоритму было предложено разбывать последовательность (6) на заданное «экспертом» число групп с последующим формированием из них латентных признаков. Для демонстрации такого варианта предложено использовать коллекцию документов «химия + технология» (K_1) и K_2 – все остальные из 10 предметных областей.

Набор из 120 слов (обобщённый словарь), упорядоченных по устойчивости (13), был разбит на 6 групп с целью формирования из них латентных признаков. Результаты кластерного анализа алгоритмом *k–means* по набору *Y*(6) показаны в табл. 7.

Таблица 7.

Контентная аутентичность при кластерном анализе

Число групп	Число латентных признаков в наборе				
	2	3	4	5	6
2	0.8765	0.8680	0.8680	0.8680	0.8680
6	0.8765	0.8810	0.8810	0.8810	0.8810
19	0.8765	0.8810	0.8810	0.8810	0.8810
25	0.8765	**0.9026**	**0.9026**	**0.9026**	**0.9026**
37	0.8771	**0.9039**	**0.9039**	**0.9039**	**0.9039**

Как видно из табл. 7 при разном числе групп значения контентной аутентичности (14) очень близки. Эта близость объясняется высокими значениями меры компактности (7) на первых 3–х латентных признаках.

Таблица 8.

Последовательность из слов относительно класса K_1 («химия + технология»)

№	Последовательность слов	№	Последовательность слов
1	Нуклеофил	9	Электрод
2	Юридик	10	Адаптация
3	Бадиий	11	Гибрид
4	Фосфор	12	дигидрохиназолин
5	Потенциал	13	Препарат
6	Матрица	14	Реабилитация
7	Фрагмент	15	Магнит
8	Аналитик	16	Гипертензия

Последовательность слов по темам из K_1 (см. табл. 8) получена из объединения словарей по 12 предметам и в определённой степени отражает специфику научных работ, разделённых на два непересекающихся класса.

В §3.2 при формировании общего словаря терминов по темам из класса «химия + технология» и последовательностей по отношению связанности получено 3 непересекающихся кластера. Для каждого кластера проведено упорядочение (ранжирование) терминов по частоте их встречаемости, результаты которого показаны в табл. 9.

Так как задача построения тематической модели из T тем является некорректно поставленной, то для её решения необходимо применять регуляризацию на основе положений академика Тихонова. В §3.3 регуляризаторы в математической лингвистике рассматриваются как дополнительные критерии оптимальности, учитывающие специфические требования решаемой задачи. Смысл интерпретируемости тем с помощью регуляризаторов заключается в следующем.

5. Сглаживание фоновых тем $B \subset T$.
6. Разреживание предметных тем $S=T\backslash B$.
7. Декоррелирование для повышения различимости тем.
8. Сглаживание + разреживание + декоррелирование для улучшения интерпретируемости тем.

Регуляризаторы для учета дополнительной информации представлены следующим образом.

5. Темпоральными моделями с модальностью времени.
6. Линейной моделью регрессии.
7. Моделями сочетаемости слов.
8. Связью родительских тем с дочерными подтемами.

Примеры комбинирование регуляризаторов в прикладных задачах:

- выявления этнорелевантного дискурса в социальных сетях;
- тематический поиск научных и научно-популярных статей;
- выявление и прослеживание событий в новостном потоке.

В качестве регуляризатора предложен параметр для сжатия пространства терминов, используемый для избежания переобучения машинных алгоритмов из-за проклятия размерности.

Таблица 9.

Ранги терминов по кластерам

Кластер		
1	2	3
илмий 5.1872	илмий 3.3079	кон 2.5673
диссерта 3.578	диссерта 2.4592	илмий 2.3278
кон 3.2724	кон 2.2433	технологи 1.983
таълим 2.5115	технологи 1.4579	диссерта 1.8233
фойда 2.0863	фойда 1.2714	фойда 0.9276
давлат 1.9492	модел 1.053	модел 0.8049
ҳуқуқ 1.9125	доктор 0.9065	энергия 0.7321
иқтисод 1.862	спектр 0.8578	техника 0.7202
технологи 1.4971	баҳо 0.7533	доктор 0.6756
баҳо 1.4962	институт 0.741	ҳарорат 0.566
доктор 1.4407	ҳарорат 0.6385	институт 0.5373
педагог 1.3692	функция 0.6312	тезли 0.5249
ўқит 1.021	тенглама 0.6297	конструк 0.5164
қонун 1.0025	электрон 0.5968	динамик 0.4661
институт 0.9021	давлат 0.578	коэффициент 0.4268
молия 0.9017	фото 0.5705	механик 0.4235
механизм 0.7699	оператор 0.5497	машина 0.4136
реферат 0.7503	энергия 0.5421	полимер 0.3935
модел 0.7259	таълим 0.5044	тенглама 0.3891
миллий 0.6924	атом 0.5033	ресурс 0.3875
инвест 0.6905	реферат 0.4997	реферат 0.3651
комплекс 0.6213	ўтказгич 0.449	металл 0.3448
статистик 0.6021	кислота 0.4358	давлат 0.3289
фоиз 0.5785	иқтисод 0.4121	масса 0.3184
ресурс 0.5756	комплекс 0.4111	баҳо 0.3165
малака 0.573	молекул 0.3875	оксид 0.3113
назарий 0.5667	техника 0.3822	кислота 0.3002
пул 0.5472	ядро 0.3822	формация 0.2868
бозор 0.5389	маъдан 0.3805	деформация 0.2743
солиқ 0.5327	ресурс 0.3761	назарий 0.2681
иммун 0.5283	коэффициент 0.3643	профессор 0.2525
тарбия 0.5028	бирик 0.3569	иқтисод 0.2391
функция 0.4972	металл 0.3487	спектр 0.236
мадани 0.4964	динамик 0.3479	синте 0.2318

Параметры регуляризации также представлены через количество латентных признаков и показатель контентной аутентичности тем.

Разработанные регуляризаторы не имеют вероятностной интерпретации. Для исследования используются отношения между документами по определяемому набору латентных признаков, которые далее рассматриваются в расширенном (но ограниченным) пространстве из сырых признаков.

Ограничение на количество сырых признаков выражается в мощности общего словаря для документов из двух непересекающихся классов. Ключевым понятием является отношение связанности объектов, которое адаптировано под задачу выбора оптимального (в локальном смысле) числа тем по критерию аутентичности.

ЗАКЛЮЧЕНИЕ

Результаты диссертационного исследования на тему «Разработка тематических классификаторов документов на узбекском языке» сводится к следующему.

1. Предложена формула для вычисления семантической связанности документов из разных предметных областей.
2. Создана технология формирования латентного пространства из терминов естественного языка. Это пространство использовалось в метрических алгоритмах кластеризации с целью разбиения документов на оптимальное количество тем.
3. Разработана методика вычисления оценок контентной аутентичности документов. В рамки методики предложен выбор оптимального разбиения описаний документов на кластеры (темы) по критерию аутентичности. Методика позволят обосновать выбор слов в предметно-ориентированные словари, формировать базы знаний с учётом омонимии и полисемии слов.
4. Разработана методика формирования общих словарей для двух коллекций документов. Методика основана на вычислении и упорядочении значений устойчивости признаков (частот терминов) для последующего включения их в словарь.
5. Получены последовательности из семантически связанных терминов из коллекций документов по предметам «физика + математика» и «химия + технология».
6. Разработана технология оценки компактности классов документов по отношению их связанности. Отношение связанности документов послужило средством для анализа кластерной структуры документов и источником новых знаний при тематическом моделировании.
7. Определен метод вычисления степени семантической связанности классов документов. Идея вычисления основана на проверке гипотезы о том, что при объединении двух документов из 2-х классов в один, его мера компактности по отношению связанности будет больше чем на каждом из них в отдельности.

8. Предложено решение проблемы омонимии, полисемии слов из разных предметных областей через формирование латентного признакового пространства. Переход к латентным признакам позволяет снижать размерность пространства для анализа, удалять малоинформативные термины и автоматизировать процесс формирования словарей для предметных областей.

9. Перспективы дальнейших исследований связаны с применением технологий оценки семантической связанности и создания общих словарей терминов на все стили тематических документов.

SCIENTIFIC COUNCIL AWARDING SCIENTIFIC DEGREES DSc.03/30.12.2019.FM.01.02 NATIONAL UNIVERSITY OF UZBEKISTAN

NATIONAL UNIVERSITY OF UZBEKISTAN

TULIYEV ULUGBEK YULDASHEVICH

DEVELOPING THEMATIC CLASSIFIERS OF DOCUMENTS IN UZBEK LANGUAGE

05.01.11 – Digital technologies and artificial intelligence

ABSTRACT OF DISSERTATION OF THE DOCTOR OF PHILOSOPHY (PhD) ON PHYSICAL AND MATHEMATICAL SCIENCES

Tashkent–2023

The theme of dissertation of doctor of philosophy (PhD) on physical and mathematical sciences was registered at the Supreme Attestation Commission at the Ministry of Higher Education, Science and Innovation of the Republic of Uzbekistan under number B2022.4.PhD/FM831.

Dissertation has been prepared at the National University of Uzbekistan named after Mirzo Ulugbek.

Abstract of the dissertation is posted in three languages (Uzbek, Russian, English (resume)) on the website (http://ik-fizmat.nuu.uz/) and the "Ziyonet" Information and educational portal (www.ziyonet.uz).

Scientific supervisor:	**Ignatyev Nikolay Aleksandrovich** doctor of physical and mathematical sciences, professor
Official opponents:	**Matyaqubov Alisher Samandarovich** doctor of physical and mathematical sciences, docent
	Norov Abdusaid Murodovich PhD in technical science
Leading organization:	**Urgench state university**

Defense will take place «____» _____2023 at _____ at the meeting of Scientific Council number DSc.03/30.12.2019.FM.01.02 at National University of Uzbekistan. (Address: 100174, Uzbekistan, Tashkent city, Almazar district, University str. 4, Ph.: (+99871) 227-12-24, fax: (+99871) 246-53-21, e-mail: nauka@nuu.uz).

Dissertation is possible to review in Information-resource centre at National University of Uzbekistan (is registered №____) (Address: 100174, Uzbekistan, Tashkent city, Almazar district, University str. 4, Ph.: +99871) 227-12-24).

Abstract of dissertation sent out on «____» _____2023 year
(Mailing report № _____on «____» _____2023 year)

M.M.Aripov
Chairman of Scientific Council on award of scientific degrees, D.F.M.S., Professor

Z.R.Rakhmanov
Scientific secretary of Scientific Council on award of scientific degrees, D.F.M.S.

D.T.Muhamediyeva
Chairman of Scientific Seminar under Scientific Council on award of scientific degrees, D.T.S., professor

INTRODUCTION (abstract of PhD thesis)

The aim of research work is development of data mining methods and criteria for solving topic modelling problem.

The object of the research work is development and justification of semantic technologies of natural language.

Scientific novelty of the research work is as follows:

proposed and substantiated the calculation of the coefficient of semantic connectedness of documents based on a quantitative assessment of the relations of document descriptions. The highest coefficient value in documents with a scientific style was obtained in the subjects of mathematics and physics;

artificial patterns from words are selected via pairs of document classes. The search for patterns was carried out in the latent feature space. Latent features were formed by disjoint groups of raw features by calculating generalized estimates;

a methodology for calculating the index of content authenticity of document collections has been developed. In the methodology has been used criteria to determine the optimal number of topics based on the results of cluster analysis;

a method has been developed to generate a common vocabulary of terms for data from two collections of documents. While including terms in the dictionary, indicators of their stability are used, calculated from the values of the class membership function.

Implementation of the research results. Based on the scientific results obtained in the development of a method for forming a common dictionary of terms for data in a collection of documents by calculating the coefficient of their semantic connectedness and the reliability indicator of the content of a collection of documents based on a quantitative assessment of their relationships when describing documents:

the results obtained in the dissertation work on creating a method for calculating the semantic relationship between sets of documents were used in the chronological analysis of text data in the grant project “Creation of the “Bakhtiyor Oila” web portal in Uzbekistan” with numbers JHBL-13. (certificate No. 10 dated June 25, 2021, Scientific Research Institute “Mahalla and Family”). The application of scientific results made it possible to form dictionaries that organize the feature space for classifying text documents;

the results obtained from the formation of a general dictionary of document collections and quantitative assessment of the relationships between document descriptions were used to create glossaries of terms in the Uzbek language and in electronic document management systems at the Scientific Research Center for Science, Technology and Marketing of the State Unitary Enterprise “UNICON.UZ” (reference No. 1 -1/300 dated February 20, 2023 State Unitary Enterprise “UNICON.UZ”). Software for developing topic models made it possible to increase the efficiency and speed of decision making by 40% by automating the process of creating dictionaries of scientific terms and electronic documents based on them.

The structure and volume of the dissertation. The dissertation work consists of the introduction, three chapters, conclusion, bibliography. The volume of the thesis is 110 pages.

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

I bo‘lim (Часть I; Part I)

1. Ignatev N.A., Tuliyev U.Y. Semantic structuring of text documents based on patterns of natural language entities // Computer Research and Modeling, 2022. Vol. 14. No. 5. – P. 1185-1197. DOI: 10.20537/2076-7633-2022-14-5-1185-1197 (01.00.00. № 3).
2. Игнатьев Н.А., Тулиев У.Ю. Анализ степени сходства и связности тематических документов на основе мер компактности // Проблемы вычислительной и прикладной математики, 2021. – № 6 (36). – С. 60-68 (01.00.00. № 9).
3. Мирзаев А.И., Тулиев У.Ю. О важности выбора первого шага при отборе информативных наборов признаков // Проблемы вычислительной и прикладной математики, 2021. – № 1 (31). – С. 118-124 (01.00.00. № 9).
4. Abdurakhmonova N., Tuliyev U. and Gatiatullin A. “Linguistic functionality of Uzbek Electron Corpus: uzbekcorpus.uz 2021” / International Conference on Information Science and Communications Technologies (ICISCT). – Tashkent, Uzbekistan, 2021. – P. 1-4. DOI: 10.1109/ICISCT52966.2021.9670043. (01.00.00. № 3).
5. Tuliyev U.Y. Space formation for the description of thematic documents // AIP Conference Proceedings, 2021. – P. 2365, 070007 // https://doi.org/10.1063/5.0056963. (01.00.00. № 3).
6. Тулиев У.Ю. Кластерный анализ текстовых документов по отношению их связности // Проблемы вычислительной и прикладной математики, 2019. – № 5 (23). – С. 102-109 (01.00.00. № 9).

II bo‘lim (Часть II; Part II)

7. Тулиев У.Ю., Лолаев М.Я. О формировании пространства для описания тематических документов // Вестник РГГУ. Серия Информатика. Информационная безопасность. Математика. – М., 2021. – № 1. – С. 35-50. DoI: 10.28995/2686-679X-2021-1-35-50.
8. Tuliyev U.Y. Intellectually analyzing documents in uzbek language / VI International conference on Computer Processing of Turkic Languages “TurkLang-2018”. – Tashkent: NAVOIY UNIVERSITETI, 2018. – P. 320.
9. Tuliyev U.Y. Classification of documents in relation to their Connectedness / ABSTRACTS of the Joint International Conference STEMM, 2019. – P. 210.
10. Тулиев У.Ю. Математическое моделирование отношения связности текстовых документов / “Matematika va amaliy matematikaning zamonaviy muammolari” mavzusidagi Respublika miqyosidagi yosh olimlar ilmiy onlayn konferensiyasi materiallari. – T., 2020. 21 may. – B. 94-98.

11. Tuliyev U.Y. Space formation for the description of thematic documents / Abstracts of the Uzbekistan-Malaysia international online conference “Computational models and technologies”. – T., 2020. August 24-25. – P. 221.

12. Tuliyev U.Y. Analysis of cluster structure of Thematic Documents / Abstracts of the International online conference “Frointier in mathematics and computer science”. – T., 2020. October 12-15. – P. 112-113.

13. Мирзаев А.И., Наврузов Э.Р., Тулиев У.Ю. Отбор информативных разнотипных признаков по правилу агломеративной иерархической группировки / “Matematik modellashtirish, hisoblash matematikasi va dasturiy ta’minot injeneriyasining dolzarb muammolari” mavzusidagi Respublika ilmiy-amaliy anjumani ma’ruzalar to‘plami. – T., 2020. 23-24 oktabr. – B. 308-310.

14. Tuliyev U.Y., Navruzov E.R., Maxarov Q.T., Raximova M.A., Mirzayev A.I. Lagranj metodi yordamida alomatlarning vaznlarini hisoblash dasturi. № DGU 10693, 09.03.2021.

15. Maxmudov M.M., Muhitdinov M.M., Tuliyev U.Y. Fan sohalarga yo‘naltirilgan lug‘atlarni ishlab chiqish avtomatlashtirilgan dasturiy ta’minoti. № DGU 22790, 20.02.2023.

16. Muxitdinov M.M., Tuliyev U.Y. Ilmiy materiallarning ko‘rsatilgan mutaxassislikka mos kelishini tekshirishning avtomatlashtirilgan dasturiy ta’minoti. № DGU 26887, 26.07.2023.

Avtoreferat O‘zbekiston Milliy universitetining «O‘zMU xabarlari» jurnali tahririyatida 2023-yil 11-noyabrda tahrirdan o‘tkazildi.

1715

Bosishga ruxsat e`tildi: 06.11.2023 yil
Bichimi 60x84 $^{1}/_{16}$. «Times New Roman»
garniturada raqamli bosma usulda chop e`tildi.
Shartli bosma tabog‘i 2,75 Adadi 100. Buyurtma № 154

“Fan va ta’lim poligraf” MChJ босмахонасида чоп этилди.
Тошкент шаҳри, Дўрмон йўли кўчаси, 24-уй.