Халықаралық ақпараттық технологиялар университеті

ӘОЖ: 004.451.5 Қолжазба құқығында

ЧИНИБАЕВА ТОЛҒАНАЙ ТЕМІРБОЛАТҚЫЗЫ

Гетерогенді құрылымы бар деректерді басқаруға арналған үлгілер мен әдістер (Big Data)

6D070400 – Есептеу техникасы және бағдарламалық қамтамасыз ету

Философия докторы (PhD)
дәрежесін алу үшін дайындалған диссертация

Ғылыми кеңесшілер
техн. ғыл.док.,
проф. Ускенбаева Р.К.

философия докторы (PhD)
проф. Cho Young Im

Қазақстан Республикасы
Алматы, 2021

МАЗМҰНЫ

НОРМАТИВТІК СІЛТЕМЕЛЕР

Бұл диссертациялық жұмыста келесі стандарттарға сәйкес сілтемелер берілген:

"Диссертацияларды және авторефераттарды рәсімдеу бойынша нұсқаулық", ҚР БҒМ, Жоғары аттестаттау комитеті, 28 қыркүйек 2004 ж. No377- 3ж.

МЕМСТ 7.32.2001. – Ғылыми-зерттеу жұмыстары туралы есеп. Ресімдеудің ережесі мен құрылымы. – Астана, 2001.

МЕМСТ 7.1-2003. Библиографиялық жазба. Библиографиялық сипаттама. Құрастырудың жалпы талаптары мен ережелері.

ҚР СТ 34.003-2002 Ақпараттық технологиялар. Ақпараттық жүйелердің мәліметтер базасының сапа көрсеткіштерінің номенклатурасы.

ҚР СТ 34.005-2002 Ақпараттық технология. Негізгі терминдер мен анықтаулар – Алғашқы рет енгізілген.

ҚР СТ. 34.015-2002 Ақпараттық технология. Автоматтандырылған жүйелердің стандарттар жинағы. АЖ құруға арналған техникалық тапсырма - Алғашқы рет енгізілген.

ҚР СТ 34.027-2006 Ақпараттық технологиялар. Бағдарламалық құралдардың классификациясы - Алғашқы рет енгізілген.

МЕМСТ 34.321–96. Ақпараттық технологиялар. Деректер базасы бойынша стандарттар жүйесі. Деректерді басқарудың эталонды моделі. Минск, 2001.

ҚР МЕМСТ 52292-2007 Электронды ақпарат алмасу. Терминдер мен анықтаулар.

АНЫҚТАМАЛАР

Бұл диссертациялық жұмыста келесі терминдерге сәйкес анықтамалар қолданылады:

Үлкен деректер (big data) - құрымдалған және құрылымдалмаған үлкен көлемдегі деректерді өңдеу құралдары мен әдістерінің топтама амалдары

Деректерді зиялық сараптау, (data mining) - өңделмеген үлкен деректер массивіндегі айнымалылар арасындағы жасырын заңдылықтар мен байланыстарды анықтау

Ақпаратты алу - құрымдалған және құрылымдалмаған құжаттардан автоматты түрде қажетті ақпаратты алу, ақпартты іздеудің бір түрі

Интероперабелдік - екі немесе оданда көп обьектілердің үйлесілімділігі

Онтология - концептуалды сұлбаның көмегімен кейбір білім саласын формалды ұсыну. Әдетте сұлба құрамына обьектілердің релевантты класының байланысы мен тәртіптері (теорема, шектеулер) құрайтын деректің құрылымынан тұрады. Онтологияны бағдарламалау үрдісінде білім саласын ұсыну үшін қолданады

Ақпаратты басқару - деректерді құру, өзгерту, жою, іздеу мен сақтауды ұйымдастырумен байланысты үрдістер

Гетерогенді деректер - әртүрлі деректер көзінен жиналған құрылымдары әркелкі деректер

БЕЛГІЛЕУЛЕР МЕН КЫСҚАРТУЛАР

DL - Description logic
DC/dc – Dublin Core
NoSQL - not only SQL
OWL - Web Ontology Language
RDF - Resource Description Framework
RDFs - Resource Description Framework Schema
SPARQL - Protocol and RDF Query Language
XML - Extenbile Markup Language
W3C - World Wide Web Consortium
IoT - Internet of Things
HDFS - Hadoop Distributed File System
SWRC - Semantic Web for Research Communities (SWRC)
FOAF - Friend of a Friend
БҚ – бағдарламалық қамтама
АЖ – ақпараттық жүйе
АТ – ақпараттық технология
ДҚ – дерекқор
ДҚБЖ – дерекқорды басқару жүйесі
ҮД – үлкен деректер
АҒК – ақпараттық ғылыми кеңістік

КІРІСПЕ

Жұмыстың жалпы нәтижесі. Қазіргі кезеңдегі қоғам мен технологиялардың дамуы адам қызметінің жаңа бағыттарын ақпараттандырумен ғана емес, сонымен бірге құзыретті басқару шешімдерін әзірлеу үшін деректерді зерттеу және талдау технологияларын кеңінен енгізумен де байланысты.

Бұл мәселені бүкіл әлемде, атап айтқанда Қазақстанда дамытуға көп көңіл бөлінеді. Еліміздің цифрлық дамудың негізгі бағыттарын айқындайтын маңызды құжат 2017 жылғы 12 желтоқсанда қабылданған «Цифрлық Қазақстан» мемлекеттік бағдарламасы [1] дәлел. Осы жобаның төлқұжатында мәліметтер көлемінің едәуір артуына байланысты үлкен деректі талдау технологиялық орталығын құрылып, олармен ары қарай жұмыста мемлекет тарапынан жәрдем болады делінген.

Өңдеу үрдісін жылдамдататын жоғары өнімділікті есептеу жүйелері мәліметтерді талдау арқылы алынатын қажетті білім жеткіліксіз. Жүйе архитектурасымен сипаттамасын құрған кезде өзара үйлесілімділік мәселесі тысқара қалып қояды. Электронды ақпараттық қорды стандарттау, ақпараттық қорды біріктіруде іздеу дәлдігін және толықтығын арттыруда пайдаланатын сәйкес құралдардың келісімі әлсіз қолданылады.

Зерттеу тақырыбының өзектілігі ғылыми ұйымдардың қызметін сипаттайтын ақпараттарды бақылау мен талдау үшін қолданылатын гетерогенді құрылымы бар үлкен деректерді басқарудың үлгілері мен әдістерін ұсынумен анықталады.

Зерттеу тақырыбының ғылыми зерттелу деңгейі. Соңғы бес-он жылда үлкен деректер мәселесімен байланысты зерттеуге алыс және жақын шетел ғалымдары өз үлесін қосуда. Атап айтсақ: А.Ф. Тузовский, Л.В. Найханова, А.Н. Бездушный, В.А. Серебряков, И.С. Михайлов, Ю.А. Загорулько, Ditmar Bayer, К.И. Шахгельдян, N. Guarino, N. Noy, M. Ehrig, A. Maedche, Yuong Im Cho және отандық авторлар: А.А. Қуандықов, Р.К. Ускенбаева, Т.Г. Балова, А.А. Шарипбаев, И.Т. Утепбергенов, Р.Р. Мусабаев, У.А. Тукеев, Н.К. Мукажанов. Әртүрлі ақпарат дерекөздерінен жиналған гетерогенді құрылымы бар деректерді біріктіру мәселесінің ғылыми сарапмасының қорытындысы бойынша аталмыш пәндік аймақтағы білім жүйеленбеген, әрі үлкен деректерді басқарудағы тәсілдемесінде бірыңғай әдістің жоқтығы ғылыми зерттеудің құрылымын, мақсатын және тапсырмасын анықтады.

Зерттеу жұмысының мақсаты – гетерогенді құрылымы бар ғылыми ақпаратты басқарудың үлгілері мен әдістерін жасау.

Зерттеу жұмысында қойылған мақсатқа келесі тапсырмалар қатарын шешу арқылы қол жеткізіледі:

1. Қолданыстағы ғылыми ақпаратты басқаруға арналған әдістер мен тәсілдеріне зерттеу жұмысын жүргізу.
2. Мәтіндер коллекциясынан берілген терминдерді тақырып бойынша бөліп іздеп табу алгоритмін әзірлеу және қолдану.
3. Ғылыми білім саласының онтологиясын құрастыру алгоритмін әзірлеу.

4. Жүйе архитектурасын дамыту, қолданыстағы жүйелермен біріктіру мүмкіндігімен тиімді жұмыс істеуін қамтамасыз ету.

5. Әзірленген алгоритмдердің нәтижелерін тестілеу және тәжірибелік бағалау.

6. Дайындалған алгоритмдердің нәтижелерін бейнелеу.

Зерттеу объектісі гетерогенді құрылымы бар ғылыми мәліметтер болып табылады.

Зерттеу пәніне құжаттарға семантикалық өзара үйлесімділікті қамтамасыз ету мақсатында гетерогенді құрылымы бар деректерді басқаратын үлгілер мен әдістер жатады.

Зерттеу әдістері. Зерттеу барысында қойылған тапсырмалар табиғи тілмен мәтінді талдау, классификация және бағдарламалық инженерия әдістерімен шешілді. Нәтижелер математикалық статистика және математикалық логика аппаратымен ұсынылды.

Диссертациялық жұмыстың **ғылыми жаңалығы** ғылыми конференциялардың хабарландыруларынан терминдерді алу негізінде, сондай-ақ Интернеттегі іздеу жүйелерінен алынған ақпаратты пайдалану арқылы ғылыми білімнің жеке саласы үшін онтологияны құрудың жаңа алгоритмі әзірленді. Оны жүзеге асырудың есептеу күрделілігінің бағасы математикалық түрде дәлелденді. Жасалған алгоритмнің айрықша белгілері: білім саласындағы терминдерді автоматты түрде бөлектеу; ғылыми білімнің басқа салаларының онтологияларын құру алгоритмін өзгертусіз пайдалану мүмкіндігі; эксперттің қол еңбегін қажет етпейді. Сондай-ақ берілген тақырып мәтіндер жинақтарынан жұп терминдерді бөлу алгоритмі әзірленді. Жұп терминдерді бөлу алгоритмінің басқа классикалық алгоритмдерден айырмашылығы мәтіндерді классификация және кластеризация арқылы салыстыруда тиімділігі жоғарылығы көрсетілді. Есептеу күрделілігінің бағасы және рубрикадағы термин салмағының негізгі функциясы оған қойылатын талаптарды қанағаттандыратындығы математикалық түрде дәлелденді.

Диссертациялық жұмысты қорғауға келесі нәтижелер ұсынылады

- ғылыми ұйымның қызметінің нәтижесін сипаттауда пәндік аймақ зерттеуінің нәтижесі негізінде онтологияны қолдана отырып гетерогенді құрылымы бар үлкен деректермен жұмыс жасағанда қолданатын алгоритмдер әзірлемесі жасалынды;

- онтологиядағы қажет ақпаратты беретін сұраныс SPARQL тілін пайдалана отырып жасалынған жүйе сұраныстарының ресми сипаттамасы берілді;

- жасалған ғылыми білімдердің жеке саласының онтологиясын құру және мәтін жинақтамаларынан жұп–терминдерді бөлу алгоритмдерінің құрамына кіретін айдардағы теминдердің салмағы базалық функциясы қойылған талаптарды қанағаттандырылатындығы дәлелденді;

- қалыптастырылған бағдарламалық қамтаманың күрделілігіне аналитикалық баға берілді.

Жұмыстың теориялық және тәжірибелік маңыздылығы: жүргізілген ғылыми зерттеудің ғылыми жаңалығы мен тәжірибелік маңызы жоғарғы

деңгейге ие. Зерттеу барысында алынған нәтижелер гетерогенді деректерді біріктіруде, оларды ары қарай өңдеу мақсатында қолдануға арналады.

Ғылыми жетістіктің шынайылығы зерттеу барысында түрлі әдістерді қолдана отырып алынған нәтижелерді ғылыми әдебиетте берілген нәтижелермен салыстырып, теориялық есептеулер тәжірибе нәтижелерінің сәйкестігімен расталды.

Автордың жеке үлесі. Автор диссертацияда баяндалған теориялық және қолданбалы зерттеулердің негізгі ғылыми нәтижелерді және қорытындыны өзбетімен алды. Бірлескен авторлық жарияланымдардың басым бөлігі диссертациялық жұмыста қойылған тапсырмалармен байланысты және бағдарламалық жүзеге асыру мен эксперименттік зерттеу нәтижелері ізденушіге тиесілі.

Диссертациялық жұмыстың апробациясы және жарияланымдары Диссертацияның негізгі нәтижелері халықаралық ғылыми конференцияларда баяндалып талқыланды: The 14th International Conference on Control, Automation and Systems, ICCAS 2014 (Оңтүстік Корея, Бусан, 2014); The 34th Chinese Control Conference and SICE Annual Conference 2015 (CCC&SICE2015) (Қытай, Хангжоу, 2015); The 15th International Conference on Control, Automation and Systems, ICCAS 2014 (Оңтүстік Корея, Гуанджоу, 2015).

Диссертациялық жұмыс Халықаралық ақпараттық технологиялар университетінің «Компьютерлік инженерия» кафедрасының және Гачон университетінің «Компьютерлік инженерия» факультетінің (South Korea, Seoul) семинарларында талқыланды.

Диссертациялық жұмысты орындау барысанда алынған нәтижелер 12 ғылыми еңбектерде жарияланды [45-56], олардың ішінде 5 мақала Қазақстан Республикасы Білім және ғылым министрлігінің Білім және ғылым саласындағы бақылау комитеті ұсынған журналдарда, 7 мақала халықаралық ғылыми-тәжірибелік конференцияларында жарияланған. Scopus компаниясының деректер базасына кіретін халықаралық ғылыми басылымда 1 мақала жарияланған (процентиль 36%).

Диссертациялық жұмыстың құрылымы мен көлемі. Диссертация құрылымы кіріспе, төрт бөлім, қорытынды, пайдаланылған әдебиет тізімі және қосымшадан тұрады. Диссертациядың көлемі 105 бет (сурет-38, кесте-11, пайдаланған әдебиет тізімі-77, қосымша-3).

Кіріспеде пәндік сала бойынша қысқаша шолу жасалынып, саланың негізгі мәселелері келтірілді. Диссертациялық жұмыстың маңыздылығы негізделіп, мақсаты мен талаптары қалыптастырылды.

Бірінші бөлімде үлкен деректерді басқару мен гетерогенді құрылымы бар ғылыми ақпаратты басқаруда қолданылатын веб-қызметтердің заманауи жай-күйіне талдау жасалынып, артықшылықтары мен кемшіліктері қарастырылды.

Екінші бөлімде бес негізгі үлгіден тұратын ғылыми ақпаратты басқару жүйесі мен оның архитектурасына жасалған жалпы формальды үлгісі берілген. Деректерді жүктейтін үлгінің алгоритмдік негізі ретінде байланыс орнататын үлгілер пәндік аймақтағы онтология мен солардың арасында болады, сонымен қатар талап етілетін функционалдылықты алуға мүмкіндік беретін әдістер

автордың көмегімен таңдап алынған, зерттеудің сәйкес келетін бағыты бойынша деректер көзінің қорытынды нәтижесі бойынша сұраныстарды орындайтын үлгілер бар. Бұл әдістерді іске асыратын бағдарлама құралдары жасалынған жүйеде қолданылу үшін сәйкестендірілген.

Үшінші бөлімде ғылыми ақпаратты басқару жүйесінде қолданылатын екі кілттік алгоритмдердің сипаттамасы бар. Мәтіндер коллекциясынан терминдерді ерекшелеп алу алгоритмі статистикалық болып саналады және төрт қадамнан тұрады. Ғылыми білім аймағындағы онтологияны құрастыру алгоритмі ғылыми конференциялар анонсы терминді ерекшелеу үшін қолданылатын дерек ретінде қолданды. Терминдердің арасындағы қатынасты анықтау үшін Интернеттегі ізденіс жүйесі арқылы табылған ақпарат қолданылды. Алгоритм семантикалық әдіс негізінде жасалды.

Әзірленген алгоритмдердің сипаттамасы осы жұмыстың ғылыми ақпараттық қорытындысы мен есептеу жүйесіне сәйкес үлгі құрайды. Бұл алгоритмдер ғылыми білім аймағында онтологияны құрап және терминдерді ерекшелеп алу мүмкіндігі туындайтын басқа да жүйелерде қолданылуы мүмкін.

Төртінші бөлімде ғылыми білім аумағындағы онтологияның құрылуы мен берілген тақырыптық бөлімдермен бірге мәтіндер жиынтығындағы терминдерді ерекшелеу алгоритмінің тиімділігін зерттеудің нәтижесі көрсетілген. Жасалынған қорытынды нәтижесінің көрсеткіші бойынша алгоритм оған қойылатын талаптарға сай және ғылыми ақпараттың қорытындысы мен есептеу жүйесінің үлгісіне сәйкес қолданылуы мүмкін. Алгоритмдер тәжірибелік білімнің аймағындағы жеке мәселеге қатысты тәуелсіз болып келеді. Олар пәндік аймақтың онтологиясын құрып, құжаттардың ішінен терминдерді ерекшелеп алу қажеттілігі туындайтын басқа жүйелерде де қолданылуы мүмкін

Қорытындыда зерттеу жұмысында алынған негізгі нәтижелер бойынша қорытынды қалыптастырылып, олардың теориялық білім ретінде және тәжірибедегі мәні берілді.

Қосымшада ғылыми зерттеудің құрамына жатпайтын онтологияны қолданып гетерогенді деректерді біріктіруге арналған техникалық материалдар келтірілген.

1 ПӘНДІК АЙМАҚТЫ ТАЛДАУ ЖӘНЕ ЗЕРТТЕУ ТАПСЫРМАЛАРЫНЫҢ ҚОЙЫЛЫМЫ

Зерттеу жұмысының бұл бөлімінде гетерогенді құрылымы бар деректердің мәселелеріне байланысты зерттеулердің қазіргі заманғы жағдайының талдауы келтірілген. Әсіресе ғылыми ақпаратты басқаруға ерекше көңіл бөлінеді.

Зерттеу саласы бойынша жүргізілген талдаулардың нәтижелері негізінде диссертациялық жұмыстың зерттеу тапсырмалары қалыптастырылды.

1.1 Ғылыми ортада үлкен деректерді өңдеудің рөлі мен маңызы

1.1.1 Деректер ғылымын дамытудағы үлкен деректердің рөлі

Қазыргі таңда үлкен деректер технологиясын бизнестің барлық салаларында, атап айтсақ телекоммуникация, логистика, денсаулық сақтау, қаржы компанияларында, сауда және мемлекетті басқаруда жаппай қолдануда. Қалыпты жағдайда деректермен жасалатын жұмыстарға деректерді жинаудан бастап, олардан құнды аналитикалық тұжырым алуға дейінгі үрдістер кіреді.

Басты мақсат – әр түрлі ақпарат көздерінен алынған үлкен мәліметтерден ақырғы қолданушыға санаулы секундтың ішінде клиенттердің іс-әрекетіндегі өзгерістерге жылдам жауап қайтару мен әрекеттердің ең ерте сатыларын анықтау мүмкіндігін беру.

International Data Corporation (IDC) зерттеулерінің нәтижесі бойынша қазір біз үшінші технологиялық платформадамыз (сурет 1.1) [2].

IDC ұсынған платформалық парадигма бірнеше платформалардың дамуын қамтиды, олардың біріншісі - негізгі компьютерлік жүйелер. Бірінші платформа 1950 жылдардың аяғында пайда болды және оның кейбір элементтері әлі күнге дейін қолданылады.

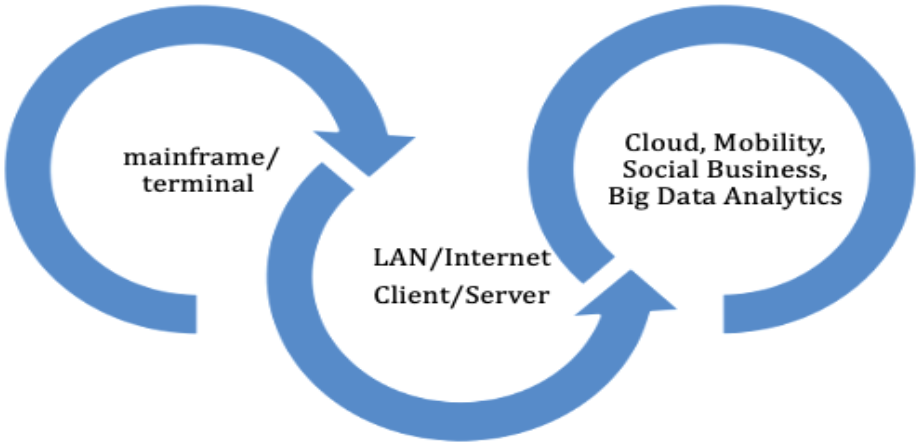
Екінші платформа 1980-ші жылдардың ортасында бастау алды. Бұл кезең қазыргі таңда әлі де қолданылатын клиент-серверлік архитектураға негізделген және де дербес компьютердің мәліметтер базасы мен мейнфреймдік қосымшаларға қосылған кезден басталды.

Үшінші платформа 2010 жылдың басында әлеуметтік интернет технологиялары, мобильді және бұлтты шешімдер, үлкен деректер технологиялар мен ішінара заттар интернетін (IoT) дамуымен байланысты.

2016 жылдың қаңтарында The Economist журналы: «үшінші платформа бұлттық онлайн-есептеумен құрылғылардың барлық түрлері өзара әрекеттесуіне негізделген [3]» деген нұсқаны ұсынды.

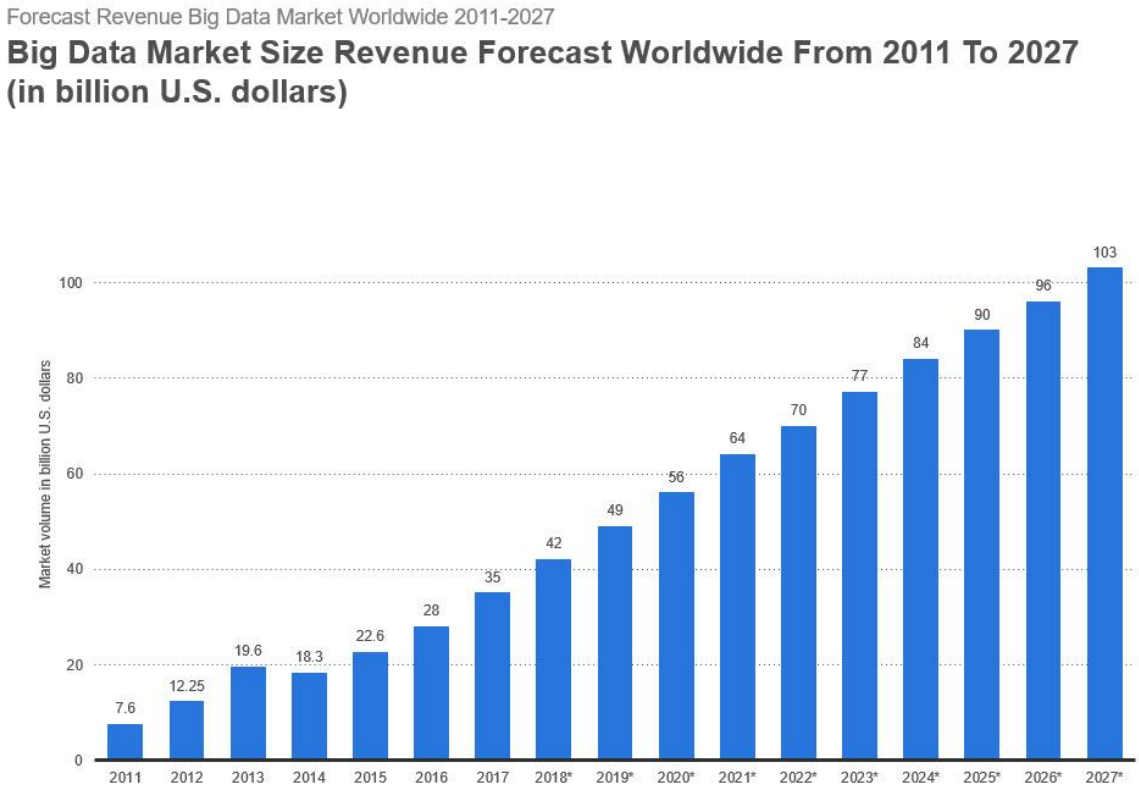
Сонымен қатар, төртінші платформа туралы кейде айтылатындығын ескерген жөн [3]. Бұл платформа интеллектуалды технологияларды, IoT-ты белсенді түрде қабылдаумен, кең таралған торлы есептеу және кванттық есептеуімен сипатталады.

Осылайша, үлкен деректерді жинау, сақтау және талдау технологиялары заманауи технологиялық платформалардың ажырамас бөлігі болып табылады деген қорытынды жасауға болады.



Сурет 1.1 – IDC үшінші технологиялық платформасы

Өз кезегінде, берілген болжамдарға сәйкес [4], әлемдік үлкен деректер нарығында бағдарламалық жасақтама мен қызметтерді сатудан түсетін кіріс 2018 жылы 42 миллиард доллардан 2027 жылы 103 миллиард долларға дейін өсіп, GAGR жылдық өсу қарқыны 10,48% жетеді (сурет 1.2) [4].



Сурет 1.2 – Үлкен деректер нарығының болжамы

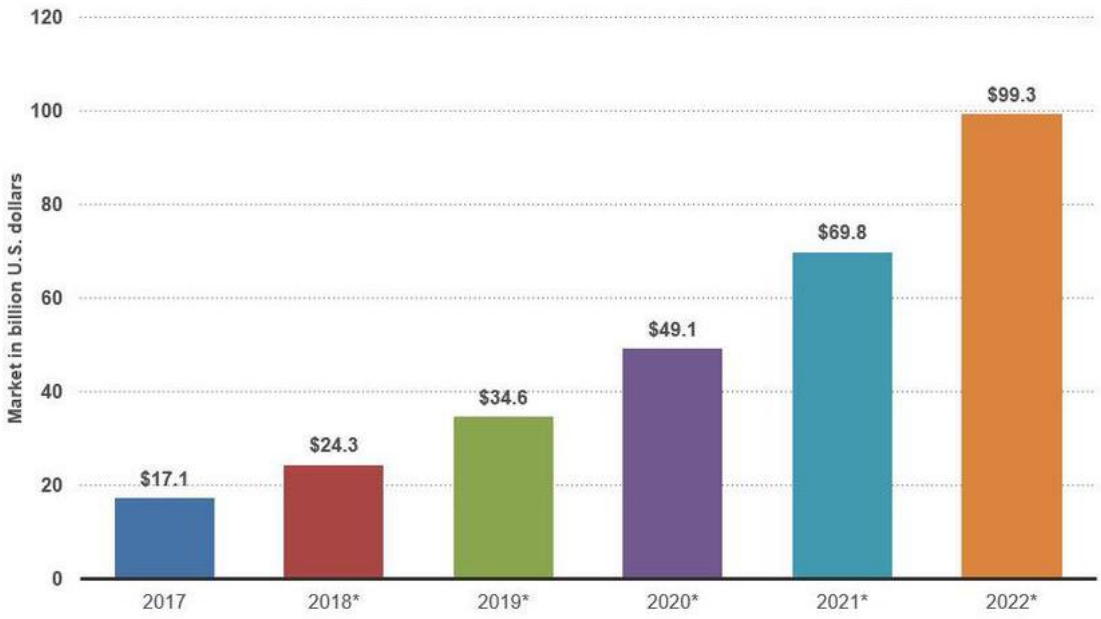
Accenture [5] зерттеуінің нәтижесінде бизнес жетекшілері 79% үлкен деректерді пайдаланбайтын компаниялар бәсекеге қабілеттілігін жоғалтады немесе жойылу алдында тұр дегенге келіссе, басшылардың 83% компанияның

нарықтағы бәсекелестік артықшылығын арттыру үшін Big Data жобаларында зерделі технологияларды іске асыру қажет деген тұжырымға келді.

Аналитика сатылымдар мен маркетинг, зерттеулер мен әзірлемелер (R&D), жеткізілім тізбегін басқару (SCM), оның ішінде тарату, жұмыс орны және операцияларды басқару, қазіргі уақытта кірістердің өсуіне үлкен үлесті қосатын алдыңғы қатарлы құрал болып табылса, аналитикалық технологиялар мен үлкен деректер зерделі сана үшін жаңа эко жүйе жасақтауға негіз болады деген тұжырымға McKinsey [6] зерттеуі келді.

NewVantage Venture Partners [5] мәліметтері бойынша, үлкен деректер шығындарды азайту (49,2%) және инновацияларға жаңа мүмкіндіктер құру (44,3%) арқылы кәсіпорындарға үлкен пайда әкеледі дейді. Кәсіпорындардың 69,4% деректерге негізделген бизнес-процестерді құру үшін үлкен деректерді қолдана бастады.

Hadoop және Big Data нарығы 2017 жылы 17,1 миллиард доллардан 2022 жылы 99,31 миллиард долларға дейін өсіп, орташа жылдық өсу қарқыны (CAGR) 28,5% жетеді деп болжануда (сурет 1.3). Жоспарланған өсудің ең үлкен кезеңі 2021 және 2022 жылдарына келеді, осы уақытта болжам бойынша нарық бір жылда 30 миллиард долларға дейін өседі [6].

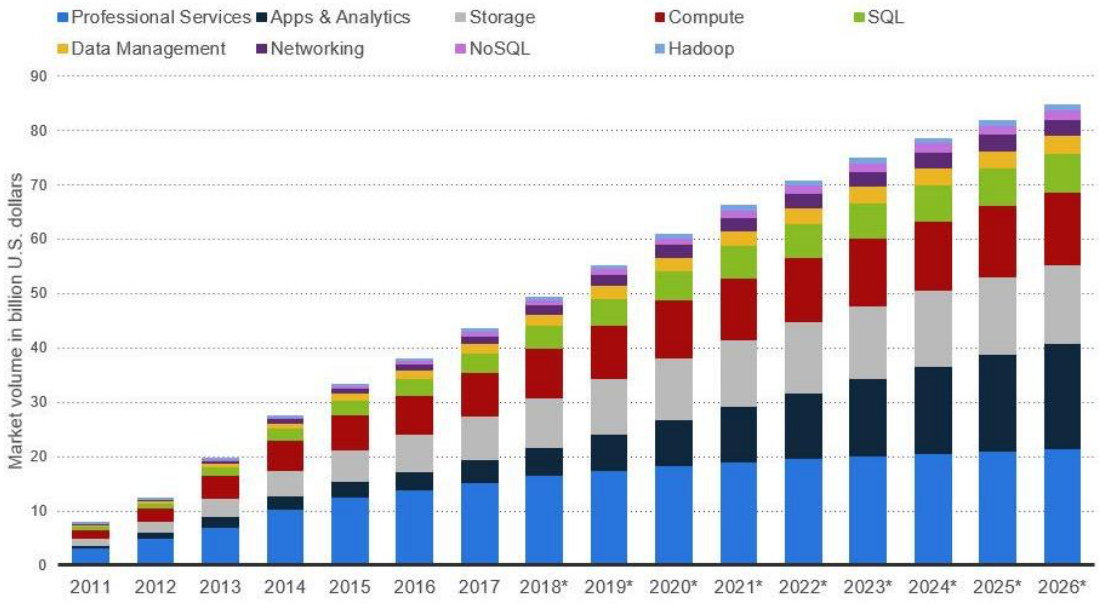


Сурет 1.3 – Нарықтағы Hadoop және Big Data көлемінің болжамы

Болжамға сәйкес [2] үлкен деректер қосымшалары мен аналитика 2018 жылы 5,3 миллиард доллардан 2026 жылы 19,4 миллиард долларға дейін өсіп, CAGR 15,49% деңгейіне жетеді. Кәсіби қызметтерді қамтитын әлемдік үлкен деректер нарығы 2018 жылы 16,5 миллиард доллардан 2026 жылы 21,3 миллиард долларға дейін өседі.

Деректерді талдаудың заманауи технологияларына және деректерге қатысты үлкен аппараттық құралдарға, қызметтерге және бағдарламалық қамтамаға әлемдік сұранысты салыстыра отырып, соңғы санаттың басымдылығы

айқындала түседі. Бағдарламалық қамтама сегменті басқа санаттарға қатысты ең жылдам өседі деп болжануда, 2018 жылы 14 миллиардтан 2027 жылы 46 миллиард долларға дейін (сурет 1.4), ал CAGR көрсеткіші 12,6% деңгейіне жетеді [2].



Сурет 1.4 - Бағдарламалық қамтама сегменттері бойынша Big Data қосымшалар нарығын болжау

Қазақстандағы мемлекеттік органдар Big Data және Open Data ашық дерек концептілерін біртіндеп ендіруде. Алғашқы қадамдардың бірі болып Павлодар қаласында орналасқан Data орталықтың құрылысы табылады. Аталмыш орталық Hewlett-Packard компаниясы мен «Қазақтелеком» АҚ бірлескен жоба. Деректер орталығы бір мезгілде электронды деректері өңдейтін 11 000 аса серверді қабылдауға қауқарлы Орта Азиядағы ең үлкен орталық.

Ақпараттық технологияларды ұлттық стандарттарға қатысты жұмыстар «Ақпаратты Қазақстан – 2020» мемлекеттік бағдарламасына сәйкес жүргізіліп жатыр.

Ұсынылған болжамды мәліметтер үлкен деректер саласындағы ғылыми және инженерлік зерттеулер тақырыбының маңыздылығын көрсетеді.

1.1.2 Үлкен деректер түсінігі мен ерекшеліктері

Википедия, 2018 жылдың ортасына қарай Үлкен деректер терминіне келесі анықтама берді: «Үлкен деректер - бұл 2000-шы жылдардың соңында пайда болған көлденең масштабталатын бағдарламалық жасақтама құралдарымен тиімді өңделетін және деректерді басқарудың дәстүрлі жүйелеріне және Business Intelligence класының шешімдеріне балама, үлкен көлемді және құрылымдалған және құрылымданбаған деректерден тұрады».

Бастапқыда 2001 ж Gartner сапаршысы Дуг Лани үлкен деректердің сипаттамалары әйгілі «3V» аббревиатурасымен біріктірілген үш сипаттама болды [6]. Алайда, үлкен деректер саласындағы соңғы зерттеулер төрт немесе

тіпті бес «Vs» туралы айтады (сурет 1.6) [7]:

- Volume - көлем;
- Velocity - жылдамдық;
- Variety - әркелкілік;
- Value - құндылық;
- Veracity - шындық.

«Көлем» термині жаппай деректерді жинау мен жинау кезінде мәліметтердің санының артуына байланысты бір компьютердің жадына жүктеп қою мүмкін еместігін білдіреді.

«Жылдамдық» термині үлкен деректердің құндылығын уақытында пайдалану үшін барлық үлкен деректерді өңдеу жылдамдығы тез болуы керектігін білдіреді.

«Әркелкілік» термині әлсіз құрылымдалған және құрылымдалмаған, мысалы, аудио, видео, веб-парақтар және мәтін, сондай-ақ дәстүрлі құрылымдалған мәліметтерден тұрады.

Төртінші термин «Құндылық» болды, бұл үлкен деректерді өңдеудің нәтижелері тұтынушыға қымбат Big Data технологияларын енгізу шығындарын ескере отырып, айтарлықтай өзгерістер әкелуді талап етеді.

Сонымен, бесінші, бірақ маңыздылығы жағынан соңғысы емес - бұл «Шынайылық». Бұл термин деректер жиынтығының қаншалықты дәл немесе шынайы болуы мүмкін екенін білдіреді. Үлкен деректер жағдайында бұл үлкен деректердің нақтылығы туралы айтқанда, тек деректердің сапасы туралы ғана емес, деректер көзі қаншалықты сенімділігін білдіреді.

Сурет 1.5 – Үлкен деректердің негізгі сипаты

Жоғарыда айтылғандарды ескере отырып, болашақта үлкен деректер «жұмыс тиімділігін арттыру, жаңа өнім жасау және бәсекеге қабілеттілікті арттыру мақсатында әрдайым жаңартылатын және әр түрлі ақпарат көздерінде орналасқан үлкен көлемді және әр түрлі құрамдағы ақпараттармен жұмыс жасауды» білдіретін технологиялар мен әдістер жиынтығын білдіреді [8].

Крейг Бэйти, Австралияның Fujitsu [8] технологиялар бойынша директоры: «Бизнес-аналитика дегеніміз - белгілі бір уақыт аралығында бизнестің қол жеткізген нәтижелерін талдаудың сипаттамалық үрдісі, ал үлкен деректерді өңдеу жылдамдығы талдаудың болжамды болуына мүмкіндік береді, болашаққа бизнес ұсыныстарын ұсына алады. Үлкен деректер технологиясы болса

гетерогенді дерек көздерінен жиналған әртүрлі дектер қоймасын талдауға мүмкіндік береді» деп есептейді.

Бизнес-аналитика мен үлкен деректердің мақсаты бір болғанымен, олар бір-бірінен ерекшеленеді:

- біріншіден, үлкен деректер іскерлік интеллектке қарағанда көбірек ақпаратты өңдеуге арналған;
- екіншіден, Big Data технологиялары жылдам шешім қабылдайтын және өзгермелі ақпаратты өңдеуге арналған, бұл дегеніміз, терең зерттеу мен интерактивтілікті білдіреді;
- үшіншіден, үлкен деректер, ғылыми зерттеудің объектісі болып табылатын құрылымданбаған деректерді өңдеуге арналған.

Тағы да бір айырмашылық, үлкен деректерде бұрын аналитикалық тапсырмаларды шешуде қолданылмаған деректер көздері болып табылады. Деректерді өңдеудің алдыңғы кезеңдерінде, әдетте, ERP, CRM, корпоративтік мәліметтер базасынан алынған ішкі деректер пайдаланылды. Үлкен деректерді ішкі дереккөздерден басқа сыртқы деректер де белсенді қолданады (сурет 1.6):

- әлеуметтік желілер
- Интернет көздері
- соңғы кезде олардың саны тұрақты өсіп келе жатқан мамандандырылған ашық деректер жиынтығы.

Сурет 1.6 – Big Data деректер көзі

Қарастырылған айырмашылықтармен қоса үлкен деректермен арналған жұмыс үрдісі өзгеше болып келеді.

1.1.3 Үлкен деректер үрдісі

Data Science үрдісі үлкен деректерге қатысты алты кезеңнен тұратын тізбекте ұсынылуы мүмкін (сурет 1.7) [9]:

1. Зерттеу мақсатын анықтау.
2. Мәліметтер жинау кезеңі.
3. Мәліметтерді дайындау кезеңі.
4. Мәліметтерді зерттеу кезеңі.
5. Мәліметтерді үлгілеу кезеңі.
6. Автоматтандыру және бейнелеу кезеңі.

Сурет 1.7 - Үлкен деректерге қолданылатын Data Science үрдісі

Бірінші кезеңде зерттеу тақырыбы, үлкен деректер технологиясын енгізуден күтілетін нәтиже, пайдаланылатынын мәліметтер, қолданылатын ресурстар және жоспарланған соңғы нәтижеден тұратын жоба тапсырмасы дайындалады.

Екінші кезеңде мәліметтер жобалау тапсырмасына сәйкес тікелей жиналады. Бұл кезеңде мәліметтердің қол жетімділігі мен сапасына талдау жасалады, деректер ішкі және/немесе сыртқы көздерден болуы мүмкін.

Екінші кезеңде алынған мәліметтер әр түрлі қателіктерді қамтуы мүмкін, сондықтан үшінші саты мәліметтердің сапасын жақсартады, оны әрі қарай пайдалануға дайындайды. Бұл кезеңде келесі фазаларды қолдануға болады:

- тазарту;
- интеграция;
- түрлендіру.

Тазарту кезеңінде алгоритмдер жарамсыз мәндерді алып тастайды және дерек көздері арасындағы деректерді сәйкестендіреді (егер гетерогенді дереккөздердің бірдей деректері әртүрлі мәнге ие болса). Интеграция кезеңінде бірнеше ақпарат көздерінен алынған ақпарат біріктіріледі. Соңында, түрлендіру кезеңі деректерді келесі қадамдарда қолдануға ыңғайлы форматқа айналдырады.

Төртінші кезең - мәліметтерді зерттеу - бұл мәліметтердің таралуы, корреляциясы, кластерлердің болуы және т.б. сияқты сипаттамаларын анықтауға бағытталған. Бұл кезеңде сипаттамалық статистика және деректерді бейнелеу әдістері белсенді қолданылады. Бұл кезең «Деректерді зерттеу талдауы» (EDA) деп аталады.

Бесінші кезеңде деректер үлгісі құрылып зерттеу мақсатының сұрақтарына тікелей жауап береді. Бұл кезең көбінесе қайталанатын болып табылады, онда сол немесе басқа үлгілер жиынтығы тексеріліп, олардың сипаттамалары анықталады. Бұл кезеңде статистика әдістері, операцияларды зерттеу, машиналық оқу және т.б. қолданылады.

Ақырында, соңғы, алтыншы кезеңде, тұтынушыға ұсыну үшін деректер бейнеленеді. Егер дамыған технологияны тапсырыс беруші бір уақытта талап

етпейтін болса, онда алынған мәліметтер үлгілері мен қолданбалы алгоритмдер негізінде автоматты қосымша жасалады.

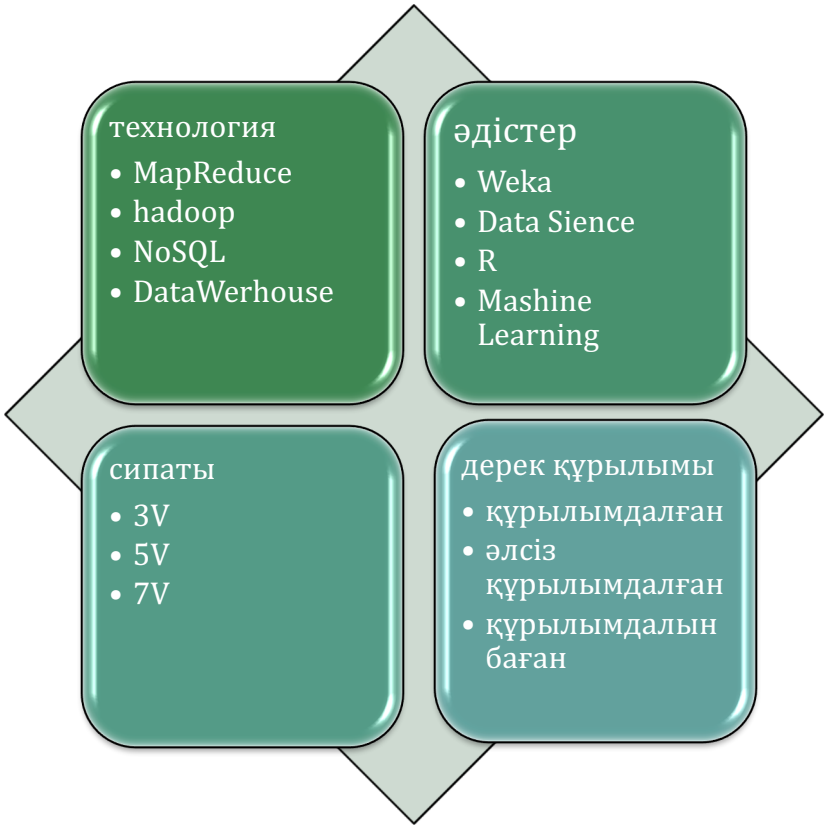
Үлкен деректерді талдаудың нақты үрдісі өзінің күрделілігі мен көпфакторлы сипатына байланысты көбінесе дәйекті түрде жүргізілмейді - зерттеуші көбінесе алдыңғы қадамдарға оралып, оларға түзетулерді енгізуге тура келеді. Сондықтан нақты үрдіс итеративті болып табылады.

1.1.4 Деректерді басқару үлгісінің технологиялары

Үлкен деректердің тапсырмасын шешудің алғашқы қадамы болып оларды бірнеше топтарға жіктеу болып табылады (сурет 1.8).

Соңғы жылдары үлкен деректерді үлестірілген сақтау мен ауқымды кластерлік жүйелерде параллель өңдеуді ұйымдастыруға мүмкіндік беретін жаңа технологиялар қалыптасты.

Үлкен деректерді кластерлік жүйеде өңдеу үрдісі келесі тапсырмалармен: деректерді процессорлар арасында бөлу және үлестіру, жүктемені теңдестіру, нәтижеден бас тарту, аралық нәтижені агрегациялау сияқты байланысты болып келеді.



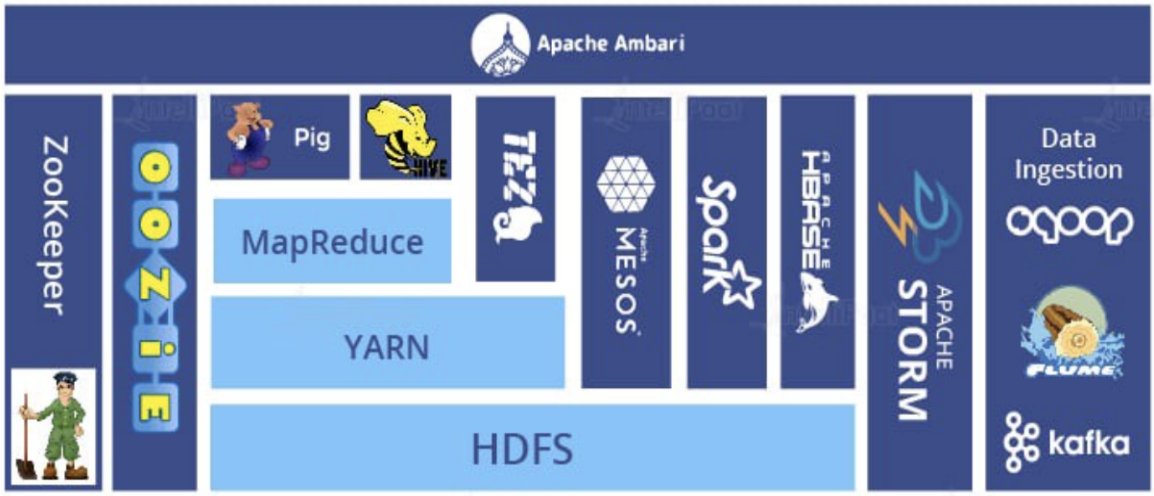
Сурет 1.8 - Big Data технологиясының негізгі сипаттамалары, технологиялары, қолдану әдістері

Мысалға, Apache Hadoop және MapReduce үлестіріп сақтау және деректерді өңдеуде қолданылады (сурет 1.9) [10]. Мұндағы ең танымал тапсырмалар мәтіндерді талдау, семантикалық іздеу және құрылымданбаған құжаттардан қосымша білім алуға байланысты болады. Бұл жерде тек қана өңделетін

мәліметтер емес, іздеу жүйелерінде жұмыс істейтін мәліметтердің үлкен массивтері болатындығын ескеру қажет. Hadoop үлестірілген фреймворк бола отырып, деректерді параллель өңдеудің артықшылығын пайдаланатын қосымшалар жасауға мүмкіндік береді.

Apache Hadoop шеңберінде MapReduce кодының ашықтығының арқасында Linux бұлтты виртуалды серверде бұлтты есептеу шеңберінде құруға болады.

Сонымен қатар Hadoop-ты кез-келген виртуалды машина мүмкіндігімен шектеліп қоймай, сонымен қатар жүйені стационарлық Linux операциялық жүйе ақаулық деңгейі төмен физикалық машиналарда орналастыруға болады [10,p. 13].



Сурет 1.9 - Apache Hadoop платформасы

Деректердің үлкен көлеміне байланысты мәселені шешу үшін NoSQL мәліметтер базасының ерекше түр-түрі жасалды [6]. Реляциялық мәліметтер базасы мен NoSQL дерекқорларының қасиеттері төмендегі кестеде келтірілген.

Кесте 1.1 - NoSQL және реляциялық дерекқор жүйелерінің ерекшеліктері

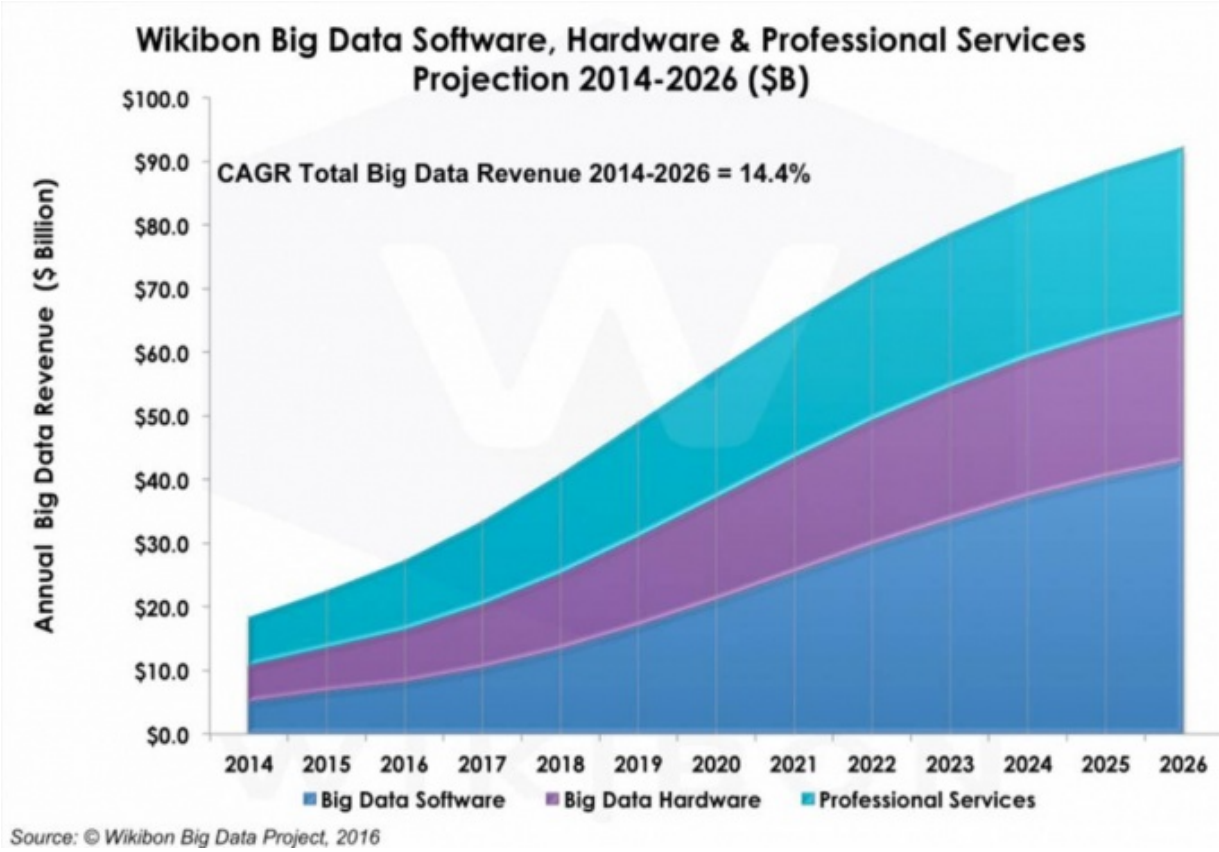
Реляциялық дерекқор	NoSQL дерекқор
Деректердің күрделі қатынастары	Қарапайым қатынас
Схемалылық	Құрылымданбаған деректердің ерікті схемасы
Ауқымдылық	Үлестірілген өңдеу
Статикалық жады	Жады есептеу қорымен бірге ауқымдалады
Әмбебап қасиеті мен функцисы	Жүйе әзірлеуші мен қосымшаға бағытталған

Өз кезегінде, біршама еңбек шығындарына қарамастан үлкен ірі кластерлерді және басқаларды дамытуда Big Data технологиясын қолдану арқылы деректердің дәйектілігі қамтамасыз етіледі.

NoSQL технологиясы (мысалы, Cassandra) реляциялық дерекқорды

ауыстыруға арналмаған, бірақ көбірек деректер қажет болған кезде мәселелерді шешуге көмектеседі. NoSQL көбінесе арзан стандартты серверлердің кластерлерін қолданады. Бұл шешім секундына гигабайттың құнын бірнеше есе азайтуға мүмкіндік береді [11].

Google корпорациясының жалпыға ортақ веб-қосымшасы болып табылатын Google іздеуіне негізделген Google Trends көмегімен «Үлкен деректер» және «ДҚБЖ» сұраныстарының сұранысын салыстыру жүргізілді. Нәтижесінде аталмыш терминдердің әлемнің әртүрлі аймақтарындағы және әр түрлі тілдердегі іздеу сұрауларының жалпы көлеміне қатысты «Танымалдық динамикасы» графигі алынды (сурет 1.10).



Сурет 1.10 - Үлкен деректер технологиясының танымалдық серпіні

1.2 Ғылыми ақпараттарды басқарудың әдістері мен тәсілдері

Мекемеде ғылыми жұмыстарды бақылау жүйесін құру бойынша жұмыс басталған кезде, барлық есеп беру қызметі ғылыми -техникалық кеңес әзірлеген қағаз бланкілер (бос кестелер) арқылы жүргізілді. Бұл үлкен көлемдегі жаңартылған ақпаратты тез өңдеуге мүмкіндік беретін қазіргі заманғы ақпараттық технологиялар қолданбағандықтан қолайсыздық туғызады. Аталған ғылыми жұмыс автомантандырылмаған, жүйеленбеген.

Зерттеу қызметінің тиімділігі мен бағытының сенімді көрсеткіштерін алу үшін жеке құрылымдық бөлімшелер үшін дербес және жалпыланған деректерді тез жинауға және сапалы талдауға мүмкіндік беретін құралдың қажеттілігі туындайды.

Қазіргі таңда қолданылатын ғылыми ақпаратты басқару әдістері мен құралдары кесте 1.2 келтірілген.

Кесте 1.2 – Ғылыми ақпараттарды басқарудың құралдары мен әдістері

Ғылыми ақпараттарды басқару әдісі	Артықшылығы	Кемшілігі
Ғылыми жұмыс есебінің ақпараты нәтижесі бойынша сандық қорытынды	қол есебі арқылы жүргізіліп, сандық көрсеткішті жазады	жұмыстың сапасы тыс қалып қояды
Конференция мен журнал материалдардың эксперттік қорытынды	эксперт сапалы, әрі мағыналы жұмыс жүргізеді	мақалалар әртүрлі тілде басылады, жіктелмеген
Негізгі мақалалардың сапартамасы	аннатациялау арқылы жұмыс көлемі азайтылады	андатпаға кірмеген мағлұматтар тыс қалып қалады
Кілт сөздер бойынша іздеу	электронды нұсқада мәтіндік ақпараттардың үлкен көлемі сараланады	фактологиялық сұраныс қамтылдмайды

Ұқсас мәселелерді шешу үшін пайдаланылатын және Интернетте ұсынылған ақпараттық жүйелердің талдауы жүйелердің бірнеше тобын анықтауға мүмкіндік берді, олардың көпшілігі библиографиялық және дерексіз мәліметтер базасы, атап айтқанда Web of Science, Scopus, Google Scholar, ресейлік портал eLibrary.ru. Олар бір немесе екі дәрежеде индекстеу және зерттеу жұмыстары сияқты функцияларды біріктіреді. Жүйелердің бір бөлігі, мысалы М.В. Ломоносов атындағы Мәскеу мемлекеттік университетінің «ИСТИНА» [12] жүйесі, Астрахан мемлекеттік университетінің «Ғылыми қызмет нәтижелері» ақпараттық -аналитикалық жүйесі [13], Elsevier-дің PURE жүйесі [14] ұйымның ғылыми қызметі мен өнімділігін бағалауға кеңінен мониторинг жүргізеді. Әлемдегі ғылыми ақпараттарды басқаруда қолданылатын ірі веб қызметтерінің сипаттамаларының салыстырмасы кесте 1.3 берілген.

Кесте 1.3 – Ірі веб қызметтерінің сипаттамаларының салыстырмасы

	№	Атауы	Құзіретті орган	Артықшылығы	Кемшілігі	Деректер форматы
Ірі веб қызмет	1	Web of Science	Thomson Scientic	Жүйеде 1900 жж бастап мақалалар тіркелген	Сұраныс кілт сөзбен ғана орындалады	.TXT
	2	Scopus	Elsever	Жалпы пәндік аумақты қамтиды	Сұраныс кілт сөзбен ғана орындалады	.TXT
	3	Google Scholar	Google	Баспаға қабылданған, бірақ жарыққа шығып үлгермеген мақалалар есепке алынады	сапасыз және жалған ғылыми жарияланымдар кездеседі	.TXT
Шет елдік проект	1	Bibster	Карлсруэ университеті, Амстердам университеті, Dresden Bank	Жүйеден деректерді RDF форматында шығарады	Жүйеге ақпарат құрылымдалған файл арқылы жүктеледі	BibTeX
	2	JeromeDL	Гданьска (Польша) Технологиялық университет, DERI (Ирландия) сандық технологияларын зерттеу Институты	Жүйе қордағы электронды ақпаратты жіктей және мазмұндай алады	Жүйеге ақпарат құрылымдалған күйде немесе қолмен енгізіледі. күрделі сұраныстар орындалмайды, ақпарат қолмен енгізіледі	BibTeX, Marc21, Dublin Core
	3	Flink	Амстердам университеті	Кілт сөз негізінде ғалымдардың интерфейстік аумағын анықтайды	Кажет пәндік аумақтағы онтологияны қолмен құрастырады	FOAF, SWRC
	4	AIR	Вульверхэмтон (Ұлыбритания) мен Аликанте (Испания) университеті	Жүйе DC құрылымдағы ақпаратты веб парақшаларынан ақпаратты жинайды	Пән аймағын үлгілейтін күрделі онтология жоқ	Dublin Core
СДҚ		Семантикалық дерекқор	Open Sourse	Кез-келген бағдарлама жасақтаушыға қол жетімді	Логикалық байласты қамтамасыз ету	RDF(s), OWL, SPARQL
Ресейлік жүйе	1	«ИСТИНА»	Ресей, Мәскеу	Кез-келген бағдарлама жасақтаушыға қол жетімді	Логикалық байласты қамтамасыз ету	RDF(s), OWL, SPARQL
	2	Астрахан университетінің «ғылыми қызметінің нәтижелері»	Ресей, Астрахан	Кез-келген бағдарлама жасақтаушыға қол жетімді	Логикалық байласты қамтамасыз ету	RDF(s), OWL, SPARQL

Web of Science Web of Knowledge жобасының бір бөлігі болып табылады. Ақпарат институтында құрастырылған жүйе құрамында 1900 жылдан жарық көрген 12 000 конференция, 11 000 астан журналда жарияланған 40 млн-ға жуық мақалалар бар. Web of Science жүйесі ағылшын тіліндегі метадеректерді кілт сөз арқылы индекстейді. Алайда өзге фактологиялық сұранысты орындамайды.

Bibster peer-to-peer технологиясы бойынша [15] жасалған еуропалық жүйе. Карлсруэ университеті, Амстердам университеті және Dresden Bank күшімен жүзеге асқан жүйе. Пәндік аумаққа тәуелсіз библиографиялық ақпараттарды ішкі тараптан көрсету үшін Semantic Web for Research Communities онтологиясы қолданылады (SWRC) [16]. Bibster артықшылығы желідегі деректерді басқа сақтау қоймаларымен біріктіруге мүмкіндік беретін RDF [17] форматында жүйеден деректерді шығару болып табылады.

JeromeDL жүйесі Польша мемлекетінің Гданьска Технологиялық университеті мен Ирландияның Сандық технология зерттеу институтында құрастырылды. Осы жерде семантикалық технологияларды қолдана арқылы жүйеге әртүрлі ресурстарды жүктеу жолы жеңілдетілді.

Тұтынушылар онтология бөлшектерінің тапсырмаларын талап ететін семантикалық іздеу жолы мен кілт сөздер бойынша қарапайым ізденісті іске асыра алады. Іздеу барысында олардың даму сатысында берілген семантикалық жүйелер үшін сирек кездесетін логикалық қорытынды қолданылады. JeromeDL жүйесі тұтынушыларға кітапхананың ішіндегі электронды ресурстарға түсініктеме беріп және оларды жіктеуге, бағалауға мүмкіндік береді. Тұтынушылар жайлы ақпаратты сақтау үшін Friend of a Friend (FOAF) онтологиясы [19] қолданылады.

Flink жүйесі ғылыми серіктестіктер туралы ақпараттарды қорытындылау мен ерекшелеу үшін Амстердам қаласындағы университетте жасалынған. Ол семантикалық технологияларға негізделген [18,p. 716]. Жүйе мақалалардың онлайн түріндегі жиынтығы мен сілтемелердің тізімі, веб-парақшалардағы ғалымдардың ғылыми байланыстары туралы (біріккен авторлар, зерттеу жобаларына бірігіп қатысу) деректердің жалпы базасына енгізуге және оны ерекшелеуге мүмкіндік берді.

Flink жүйесінде қосымша онтология ретінде басқа да көрсеткіштер мен қарастырылатын бағыттағы тапсырмаларды шешудегі әртүрлі зерттеушілердің ғылыми үлестерін бағалу үшін FOAF және SWRC сөздіктері қолданылады. Бұл жүйенің мақсаты ғылыми серіктестіктегі адамдардың әлеуметтік байланыстары туралы ақпараттардың қорытындысы мен жиынтық тәсілін жасаудан тұрады, алайда қосымша факторлармен бірге онтологияның автоматты түрде құрау модулінің болмауы қызығушылық танытқан ғалымдардың тізімін басынан ұсынуды талап ету, ішкі деректердің ерекшеліктеріне байланысты ғылыми ақпараттарды қорытындылау үшін Flink жүйесі қолдануға келмейді.

AIR жүйесі Вулверхэмптон (Ұлыбритания) қаласы мен Аликанте (Испания) қаласындағы университеттердегі ғалымдардың біріккен жобасының аумағында жасалынған [20]. AIR мақсаты университеттік веб-сайттарға енгізілген ақпараттар туралы университет қызметкерлерінің мақалаларын индекстеу болып табылады. Алдымен деректер AIR-ға автоматты түрде жиналады, содан кейін

порталға енеді, редакцияланады және жиналған деректердің дұрыстығын бекітеді. AIR жиынтығы мақалалардың толық мәтінін іздемейді, тек қана әріптесінің мақалаларының тізімі сияқты библиографиялық тізімнен тұратын веб-парақшаларды анықтайды.

AIR жиынтығында күрделі сұраныстарды іске асырылмайды, енетін ақпараттарды қорытындылайтын қалыптасушы пәндік аймақ және күрделі онтологиялардың жоқтығы кемшілік болып табылады. Метадеректердің ерекшеленуі ғылыми жұмыстың онотологиялық деңгейінде орын алады. Іздеу интерфейсінің болмауы алынған деректердің тақырыптық қорытындысын жүргізуге мүмкіндік бермейді.

Жоғарыда аталған артықшылықтар ғылыми ақпаратты жаһандандыруға септігін тигізгенмен соңғы қолданушының сұранысын толықтай қамтамасыз етпейді [21]. Сондықтан осы диссертациялық жұмыста ғылыми қызметкердің жұмысын көрсететін автоматтандырылған кешеннің нобайын жасау ұйғарылды.

1.3 Зерттеу жұмысына қойылатын тапсырмалар, зерттеудің әдістері мен деңгейлері

Ғылыми ұйымдарда ғылыми қызметкердің жылдық есебінде сандық көрсеткіштері көбіне сапасыз болып келеді, кейде мүлдем ескерілмей де қалады. Әлемдік деңгейге шығу үшін ғалымның паттентік зерттеулері, конференциядағы баяндамалары мен академиялық жұмыстары (дипломдық және диссертациялық жұмыстарына басшылығы) сарапшының қатысуынсыз есепке алынып мазұнына анағұрлым көңіл бөлінуі қажет.

Берілген диссертациялық жұмыстың шеңберінде отандық ғылыми кызметкердің жұмысын басқаруды оңтайландыру мақсатында жарияланым деректерді жылдам жинауға және талдауға мүмкіндік беретін интеллектуалды ақпараттық-аналитикалық жүйе құралын құру қажеттілігі туындады.

Жүйе келесі шарттарды қанағаттандыру керек:

- қауіпсіздік (пайдаланушылар мен парольдердің дамыған жүйесінен тұрады);
- кіру құқығын дифференциациялаумен көп қолданушыға қол жеткізу;
- оқытудың әр түрлі деңгейлері бар пайдаланушыларға арналған ыңғайлы және түсінікті интерфейс;
- университеттің бірыңғай ақпараттық кеңістігіне бірігу.

Бұл әзірленіп жатқан қосымша қызметі Интернетке кез келген уақытта қосылғанда қол жетімді етеді. Жүйені әзірлеушілер тобының алдында тұрған негізгі міндеттердің ішінде мыналарды бөліп көрсетуге болады:

- университетте қолданылатын ғылыми жұмыстар бойынша есеп беру формаларын талдау;
- ақпараттық жүйе үшін мәліметтер қорын және оның объектілері арасындағы байланыстарды әзірлеу мен үлгілеу;
- интерфейске эргономикалық талаптарды ескере отырып экрандық формаларды әзірлеу;
- форма өңдеушілерге арналған бағдарлама кодын әзірлеу;

- ақпараттық жүйенің серверлік бөлігінің бағдарламалық кодын әзірлеу. Әзірленген бағдарламалық қосымша ұқсас бағдарламалық өнімдерге қойылатын барлық заманауи талаптарды ескеруі және заманауи технологияларды қолдана отырып жасалуы тиіс [22].

1 бөлім бойынша тұжырым

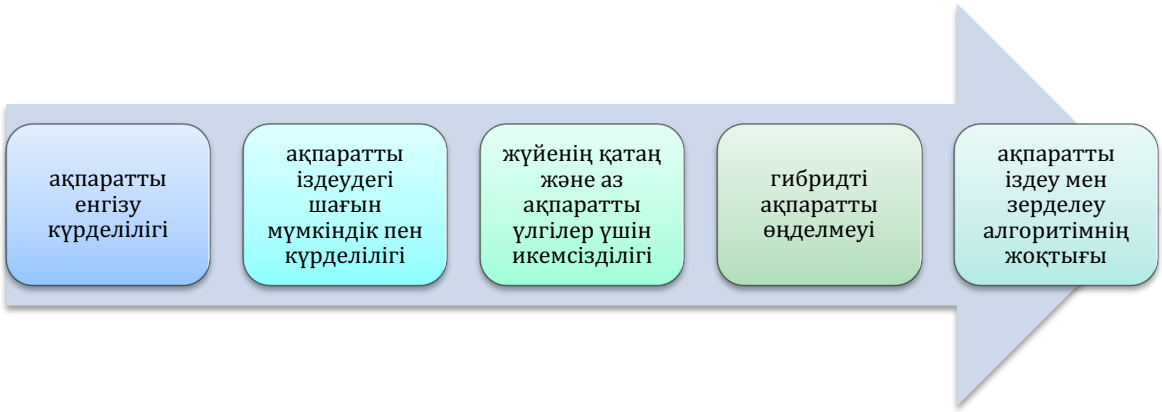
Диссертацияның алғашқы бөлімінде ғылыми ақпаратты басқару мәселесі кеңінен қарастырылған.

Негізгі тапсырманың мүмкін болатын шешімі ретінде қарастыруға болатын ғылыми деректердің қорытындысы мен өңделу жүйесінің қазіргі уақытта белгілі негізгі кемшіліктерін тізіп, атау үшін ғылыми және тәжірибелік әдебиеттер көп қарастырылған.

Қорытындыда келесі екі жағдайды атап өту маңызды.

1. Жасалынған бағдарламалық жиынтық жеке компоненттер мен ақпараттық белсенділерді жеке қолдану, жоғарыда көрсетілген басқа да жүйелермен байланысу мүмкіндігіне ие.

2. Бірінші бөлімде пәндік аймақ ретінде ғылыми ақпаратты басқару жүйесі мен тәсілдері, әдістерін зерттеу туралы айтылған, бұл осы аймақтағы мәселені жүйелі түрде көруге мүмкіндік береді. Бұл мынандай нәтиже шығарады:



Сурет 1.11 – Ғылыми ақпаратты басқару жүйесінің автоматтандырылған жиынтықтың жалпы формальды үлгісі

Ғылыми жұмыс туралы деректердің қорытындысы мен жүйелендірілуі, сақталуы, іздеу аймағындағы автоматтандырушы технологиялық үрдіс жиынтығын жасауға жарамды жағдай туғызу;

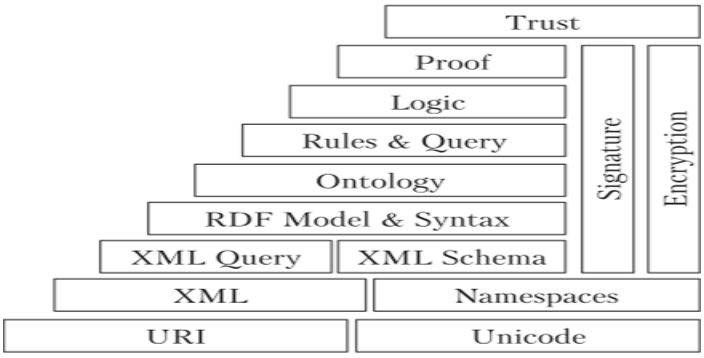
Оны құрайтын жеке бөлшектердің аса нақты үлгілер мен осындай автоматтандырылған жиынтықтың жалпы формальды үлгісін құру қажет.

2 ҒЫЛЫМИ АҚПАРАТТЫ ЕСЕПТЕУ, ТАЛДАУ ЖҮЙЕСІНІҢ АРХИТЕКТУРАСЫ

Берілген бөлімде ғылыми ақпаратты автоматты басқару жүйесі кезінде қолданылған архитектура-технологиялық шешімдер көрсетілген.

2.1 Semantic Web технологиясы

Семантикалық веб – ақпараттың машиналық өңдеуге жарамды түрде ұсынылуын стандарттау арқылы World Wide Web негізінде құрылған жалпыға қолжетімді жаһандық семантикалық желі. Semantic Web технологиялардың құрамы төменде келтірілген (сурет 2.1):



Сурет 2.1 – Semantic Web технологиясының элементтері

Ары қарай, берілген бөлімде онтология, метадеректер мен логикалық қорытындылар толығымен түсіндіріледі.

2.1.1 Онтология

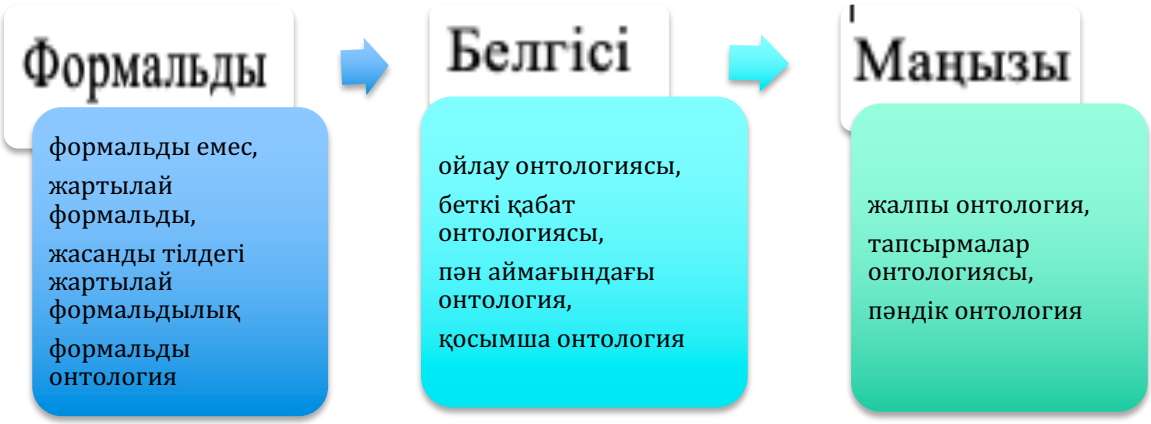
Онтологияны жасанды интеллект аймағында жұмыс істейтін зерттеушілер 1980 жылдардан бастап информатика аймағында қолдана бастаған. Алдымен олар жасанды тілдерді өңдеу үшін, кейін білім түсінігін көрсетуде қолдана басталды. 1990 жылдары онтологияны Интернет желісі мен дерекқордан ақпаратты іздеу үшін қолдану мүмкіндігіне қатысты зерттеу бастады. Кейінірек онтология семантикалық веб-желінің концепциясын іске асыру үшін қолданылатын негізгі элементке айналды.

Онтология – бұл нақты сипаттама, машина арқылы оқылатын құрал. Қазіргі уақытта онтология формальдылық деңгейі бойынша, пайда болғызу мақсаты мен мазмұны бойынша деген 3 тәсілді қолданып жіктелінеді.

- *формальдылық деңгейі* – формальды емес, жартылай формальды, жасанды тілдегі жартылай формальдылық, формальды онтологиядан тұрады.

Онтолгияны жіктеу шеңбері бойынша 4 деңгейді ерекшелеп айтуға болады: ойлау онтологиясы, беткі қабат онтологиясы (мысалы, Cyc, DOLCE, SUMO), пән аймағындағы онтология мен қосымша онтология.

- Классификацияның маңызы бағыт бойынша онтологияның классифкациясына қатты ұқсас, бірақ онда абстрактілі мақсатқа емес, нақтылы мазмұнына басты назар аударылады. Маңызы бойынша онтологияның негізгі түрлері: жалпы онтология, тапсырмалар онтологиясы, пәндік онтология болып табылады (сурет 2.2).



Сурет 2.2 – Онтологияның жалпы жіктелуі

Қазіргі уақытта онтологияны сипаттау үшін бүкіл әлемдік тордың консорциумы арнайы OWL – Ontology Web Language [23] тілін қолданылады.

OWL тілі алдыңғы уақытта қолданылған OIL [34], DAML+OIL [35] және RDFS [13,p. 122] тілдермен салыстырғанда деректердің семантикасын сыйпаттауда зор мүмкіндігіне ие және де үш бөліктен тұрады:

1. OWL Lite – шектеулерді жеңілдету мен классификациялық иерархияны туғызуға арналған, ол жеңіл іске асады;

2. OWL DL (логиканы сипаттау) – толықтығы мен рұқсаттамасын сақтаған кездегі максималды түрде айтылуын қолдайды;

3. OWL Full – синтаксистік еркін RDF пен максималды түрде айтылуын қамтамасыз етеді, бірақ көрсету қабілетілігіне кепілдік болмайды.

Семантикалық технологияларда қолданылатын тілдер төмендегідей жіктеледі (сурет 2.3):

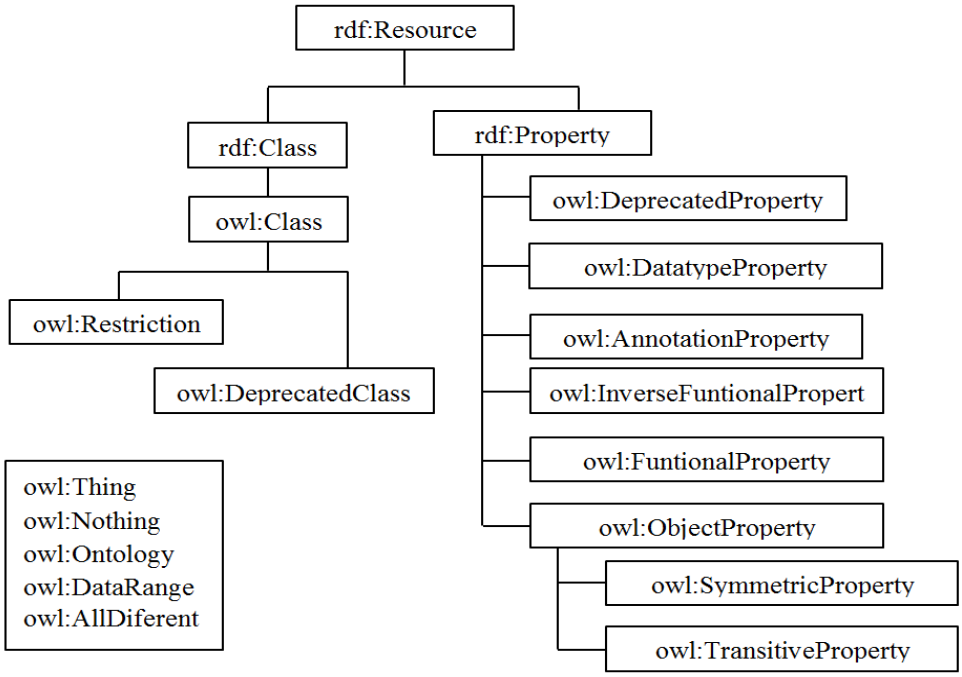
Сурет 2.3 – Тілдер классификациясы

OWL DL мен OWL Lite тілдері бағдарламалық өңдеу үшін RDFs сөздігін қолдануға шектеу қояды. Бұл шектеулер дескрипциялық логикалардың (ДЛ) негізінде OWL-онтологиясында логикалық қорытынды жасай алатын FaCT++ мен Pellet сияқты жүйелерді қолдану арқылы логикалық қорытынды жүйесіне рұқсат ету мен толықтай есептеуіне кепілдік береді [24]. Берілген тілдің басқа да қатынастары кесте 2.1 көрсетілген [24,p. 15].

Кесте 2.1 – OWL тілінің арнайы қатынастары

OWL-онтология байланысының сипаты	Сәйкес байланыстың шығару ережесі
$owl: inverseOf$	$(p_1, owl: inverseOf, p_2) \wedge (s, p_1, o) \rightarrow (o, p_2, s)$
$owl: TransitiveProperty$	$(p, rdf: type, owl: TransitiveProperty) \wedge (X, p, Y) \wedge (Y, p, Z) \rightarrow (X, p, Z)$
$owl: FunctionalProperty$	$(p, rdf: type, owl: FunctionalProperty) \wedge (s, p, o) \rightarrow (s, p, o)$ s үшін жалғыз болып келеді
$owl: InverseFunctionalProperty$	$(p, owl: InverseFunctionalProperty, p_2) \wedge (s, p_1, o) \rightarrow (o, p_1, s)$ және (o, p_1, s) бір дерекқорда дара болып келеді
$owl: sameAs$	$(a, owl: sameAs, b) \rightarrow (b, owl: sameAs, a)$

Онтологияны сипаттайтын редакторлар. Қазіргі уақытта әртүрлі форматтар мен тілдердегі онтологияны импорттайтын, қарайтын, өңдейтін, онтологияның кітапханасын басқаратын және графиктік өңдеуді қолдайтын мүмкіндіктері бар көптеген бағдарламалар кездеседі. Бұндай құралдарға мысал ретінде Ontolingua, Protégé, OntoSaurus, OntoEdit, OilEd [25], WebOnto, WebODE сияқты жүйелерді келтіруге болады.



Сурет 2.4 – Онтологияның OWL негізгі элементтерінің жалпы құрылысы

Protégé редакторларының ерекшеліктері:

- Қосымша аймақтың отнологиясын орнату үшін қолданылатын әрекеттер (жасалу, өңделу мен қарастыру). Оның бастапқы мақсаты бұл үлгілерді бағдарламалық кодқа енгізу, пәндік аймақтардың үлгісін жасау, қолдауды бағдарламалық қамтамасыз етуде әзірлеушілерге көмек көрсету;
- Нақтылы пәндік класстар немесе иерархиялық абстракт құрылысын теріс аудару арқылы онтологияны жобалауға мүмкіндік беру;
- Онтологияның құрылысы каталогтық онтологияның иерархиялық құрылысына ұқсас. Protégé құрастырылған онтологиясының негізінде класстар мен олардың бөлімдерінің көшірмелерін енгізу үшін білім алу формасын шығарады;
- Тәжірибесі жоқ қолданушылар үшін ыңғайлы интерфейсі бар;
- Сақталатын және әртүрлі форматтардағы үлгілерді өңдеу үшін оны үйретуге мүмкіндігі бар плагиндерден тұрады.

2.1.2 Мета деректер

Жалпылама түрде мета деректер дегеніміз – берілген сипаттамалардан тұратын құрамдастырылған деректер жайлы түсінік [26]. Мета деректер ақпараттық жүйенің объектісін сипаттайды және объектінің әртүрлі аспектілерін сипаттайтын элементтердің (қасиеттері мен атрибуттары) жиынтығынан тұрады.

Мета деректердің құрамы мынандай сипаттамаларға сай болады: атрибуттардың қорытынды жиынтығы; атрибуттардың атауларының бар болуы; әрбір атрибуттың бекітілген мәні (атрибут мәнінің түсіндірілуі); бір атрибутқа бірнеше мәнді беру мүмкіндігі.

Мета деректердің көмегімен сипатталатын кез-келген объект әртүрлі үш жағдайда қарастырылуы мүмкін: құрамы, мәнмәтін және контент.

Құрамы. Объект ішкі құрам сияқты, сыртқы құрам болып та сипатталады – ақпараттық жүйедегі басқа да объектілермен байланысты. Объект қаншалықты көп құрамнан тұрса – сыртқы жағынан да ішкі жағынан да, соншалықты объекті табу, басқа объектілермен сәйкестігін анықтау мен басқару жеңілдей түседі.

Мәнмәтін. Мәнмәтін қасиеттерінің объектісіне байланысты сыртқы болып табылады, және оны кім қашан, қалай әрі не үшін жасағандығы бойынша анықталады. Мәнмәтін көптеген басқа объектілердің ішінен объекті анықтауға мүмкіндік береді.

Контент. Объект қолданушыларға қажетті ақпараттарды ұсыну үшін ақпараттық жүйеде жасалынады, және бұл ақпарат объектінің ақпараттық мазмұны арқылы беріледі, яғни, ол – контент.

Мета деректер жоғарыда келтірілген үш түрдің ішінен бір немесе одан көп аспектіні сипаттай алады. Семантикалық техналогия шеңберінде объектінің мәнмәтіні мен контентіне сипаттама беретін мета деректерді зерттеуге аса көп уақыт бөледі. Берілген диссертацияда семантикалық мета деректерді анықтаудың төмендегідей жолдары қолданылады.

Семантикалық мета деректер – бұл онтологияны сипаттайтын белгілі бірнеше тілдерде болған пәндік аймақтың түсінігі арқылы ақпараттық жүйедегі объект контенті мен мәнмәтінін сипаттайтын мета деректер.

W3C мета деректерін сипаттау үшін RDF – Resource Description Framework ресурстарын сипаттайтын арнайы тіл бар.

RDF тілі. RDF әртүрлі бағдарламалардың арасында құрамдастырылған мета деректерді қайтадан қолдану мен алмастыру, кодтау қызметтерін орындауға мүмкіндік беретін машинамен оқытылатын тіл. XML мета тілінен басталып зерттелінетін бұндай мүмкіндіктер семантикалық ақпараттардың пайдасын арттырады [27]. XML тілі мета деректердің иерархиялық құрамын сипаттайды. RDF бағдары бар граф түріндегі ресурстарда болатын ақпаратарды береді.

RDF ресурстар басқа да ресурстармен байланысын сипаттау үшін үштікті қолданады. RDF-үштігі: s – субъект (URI-идентификатор, немесе бос түйін), p–предикат (URI- идентификатор), o–обьектен (URI- идентификатор, немесе литерал). RDF-үштіктерінің көпшілігі қабырғалары предикат ретінде белгіленген, шыңы субъект пен объект болып табылатын бағдары бар бағанды көрсетеді. Сурет 2.5 қарапайым сипаттама берілген.

Сурет 2.5 – RDF- сериаландырудың үштігі мен форматы

RDF-деректері әртүрлі форматтарда жазылуы мүмкін, мысалы N3, N-Triples, Turtle, RDF/XML. Олардың ішінен адам үшін ең қолайлысы (адам оқи алатыны) N3 мен Turtle форматтары болып табылады.

2.1.3 Семантикалық технологиялардағы логикалық тіл

Дескрипционды логикалар (ДЛ). Онтологияны сипаттау әртүрлі логикаларға негізделеді. Сондай логикалардың бірі дескретивті логика. Сипаттайтын тілдер: OIL, DAML+OIL, DAML-ONT, мен OWL.

OWL тіліндегі логикалық қорытынды: рұқсат ету, көрсете алу мен түсініктің автоматты классификациясында қолданылады. Бағдарламалық жүйеде логикалық қорытындыға негізделген операцияларды орындауда логикалық тілді қолдану арқылы жауап алуға болады.

ДЛ-де синтаксистік құрылыстық блок болып табылатын атомдық түсінік (бір ғана предикат), атомдық рөлдер (бинарлы предикат) және елшілер (индивидтар, константтар) бар. Ролдер (қасиет, байланыс) ары қарай түсінікпен байланысты болатын тәуелсіз элементтер болып табылады.

Түсініктер мен рөлдерді конструкторлардың көмегімен (операциялар) күрделі түсініктерді сипаттау үшін көріністерді біріктіруге болады.

Semantic Web Rule Language (SWRL) – Семантикалық тордағы ережелерді сипаттау тілі [28]. OWL-DL онтологиясының сипаттамасы онтологияны сипаттауға ережелерді қосуға мүмкіндік береді. Ол OWL мен RuleML тілдерінің бірігуіне негізделген.

2.1.4 Сұраныстардың тілі

SPARQL семантикалық деректерге қатысты сұранысты орындауда кеңінен қолданылады. SPARQL басқа тілдерден келесідей артықшылықтарға ие:

- RDF түріндегі сақталатын деректерге сұраныс тастау үшін қолданылады.
- Олардың конъюнкция және дизъюнкциялармен бірге сұраныстарды қалыптастыру мүмкіндігіне ие.
- RDF-графының көмегімен рұқсат етілген сұраныс тастаудың мәні мен шектеулерін тестілеуді қолдайды;
- SPARQL сұранысының нәтижелері RDF–графтары мен қорытындылайтын жиынтық түрінде берілуі мүмкін.

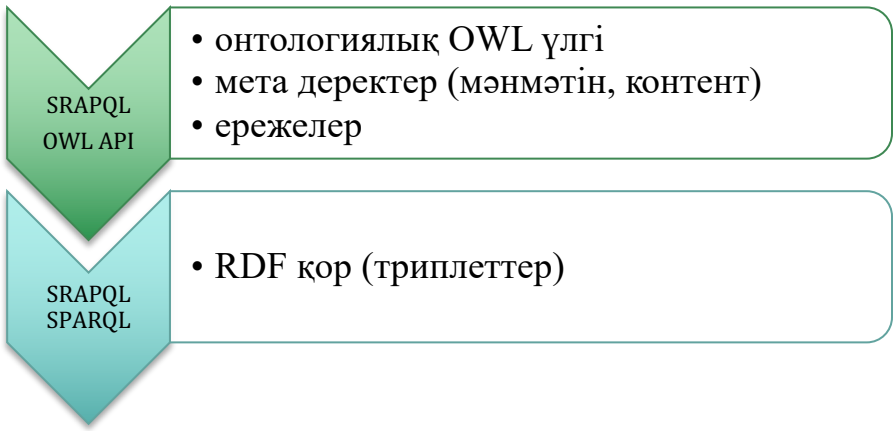
SPARQL шаблондардың негізгі графы RDF-үштігіне ұқсас болып келеді, триплет шаблондарының жиынтығынан тұрады.

2.2 Семантикалық дерекқор

2.2.1 Семантикалық дерекқор түсінігі

Дерекқор (ДҚ) деп кейбір пәндік аймақтағы, тақырыпқа, тапсырмаға қатысты құбылыстар немесе жағдайлар, үрдістер, объектілер жайлы деректердің белгілі бір тәртібімен сәйкес келетін ұйымдастырушылықты атайды. Ол қолданушылардың ақпараттық қажеттіліктерін қамтамасыз ету үшін ұйымдастырылған, сонымен қатар бұл деректердің толығымен және кез-келген бөлігін әсерлі етіп сақтау үшін қажет.

Семантикалық ДҚ (СДҚ) – онтология, семантикалық мета деректер, және логикалық ережелердің жиыны. Оны $DB_s = \{O, M, R\}$ қатынасы арқылы сипаттаймыз, мұнда O – онтология, M – семантикалық мета деректер, R – көптеген логикалық ережелер (сурет 2.6).



Сурет 2.6 – Деректердің семантикалық негізінің құрамы

Белгілі деректердің негізінде логикалық қорытындыны алу үшін R логикалық ережелері қолданылады. $R = \{R_1, R_2\}$, мұндағы R_1 – онтологиялық логикалық ережелер жиыны; R_2 – мүмкін болатын логикалық қорытындыларды алуда пайда болған қолданушылардың логикалық ережелер жиыны.

2.2.2 RDF-қоймасы

Семантикалық ДҚ RDF-қойма негізінде құралады. RDF-қоймасының (RDF-Stores, RDF-triple store) RDF-деректерге (үштігіне) қол жеткізуге және оны сақтау үшін арналған ақпараттық жүйе.

RDF-қоймасы өзінің архитектурасы бойынша екі түрге бөлінеді: шынайы RDF-қойма (native stores) мен RDF-қоймасы, оларға дерекқорды басқару жүйесі (DBMS-backed stores) көмегімен қолдау беріледі.

RDF-қоймасы (native stores) толығымен ДҚБЖ тәуелсіз жұмыс істейді және RDF деректерін өңдеу үшін оптимизацияланған ДҚ ядроны толығымен іске асырады. Бұл қоймаларда деректер файлдық жүйенің ішінде сақталынады, ол бір ғана файлда болады немесе бірнеше файлдарға бөлінеді.

Реляционды ДҚБЖ (DBMS-backed stores) арқылы қолдау берілетін RDF-қоймасы ДҚБЖ бар деректерді іздеу мен сақтау қызметтерін орындау үшін қолданылады.

Сақтаудың екі түрлі үлгісі бар: жалпы сұлба және нақтылы онтологиялық сұлбасы.

RDF-қойманың жалпы сұлбасы үштіктерді сақтау үшін әртүрлі кестелерден құралуы мүмкін. Мысалы үш бағаннан тұратын кестелер: субъект, предикат және объект. Сонымен қатар жеке кестелердегі литералдар мен URI-идентификаторлардың барлығын ДҚ сақтау жолы сияқты сұлбаның нормаландырылған принциптерін сақтау қажет. Бұндай әрекет жалпы сұлбаны қолдану арқылы сақтау сұлбаларының оңай болуын көздейді, алайда күрделі сұрақтарға жауап беру үшін кестелердің көп мөлшерін біріктіруді талап етеді.

Нақтылы онтологияның сұлбасының негізінде әрекет ету құрамы нақтылы онтологияның құрамдық қасиетін көрсететін арнайы кестелер жиынтығындағы үштіктерді сақтаумен байланысты болады. Сұлбаның үш түрін еркшелеп алуға болады:

- Онтологияның әрбір классы үшін (көлденең ойлау) жеке кесте жасалынады, оның әрбір бағаны кластың бір ғана қасиетіне сәйкес келеді, ал әрбір жол – класстың бір ғана экземплярына сәйкес. Бұндай әрекеттің негізгі кемшілігі онтологияның құрамы өзгерген кездегі ДҚ кестесін қайта құрастырылу керек;
- Әрбір предикат үшін субъект пен объекттен тұратын жеке кесте жасалынады (тігінен көрсету). Бұл ДҚ байланысты күрделі сұраныстарды орындаудың әсері айтарлықтай төмен;
- Гибридті сұлба – алдыңғы әрекеттердің екеуінің де артықшылықтарын біріктіреді, сонымен қатар әрбір класс көрсетілген кластың экземплярлар идентификаторынан тұратын кесте ретінде беріледі.

Белгілі RDF-қоймалары – 4/5store, AllegroGraph, OWLIM жатады.

2.3 Семантикалық дерекқор негізіндегі ақпараттық жүйелер

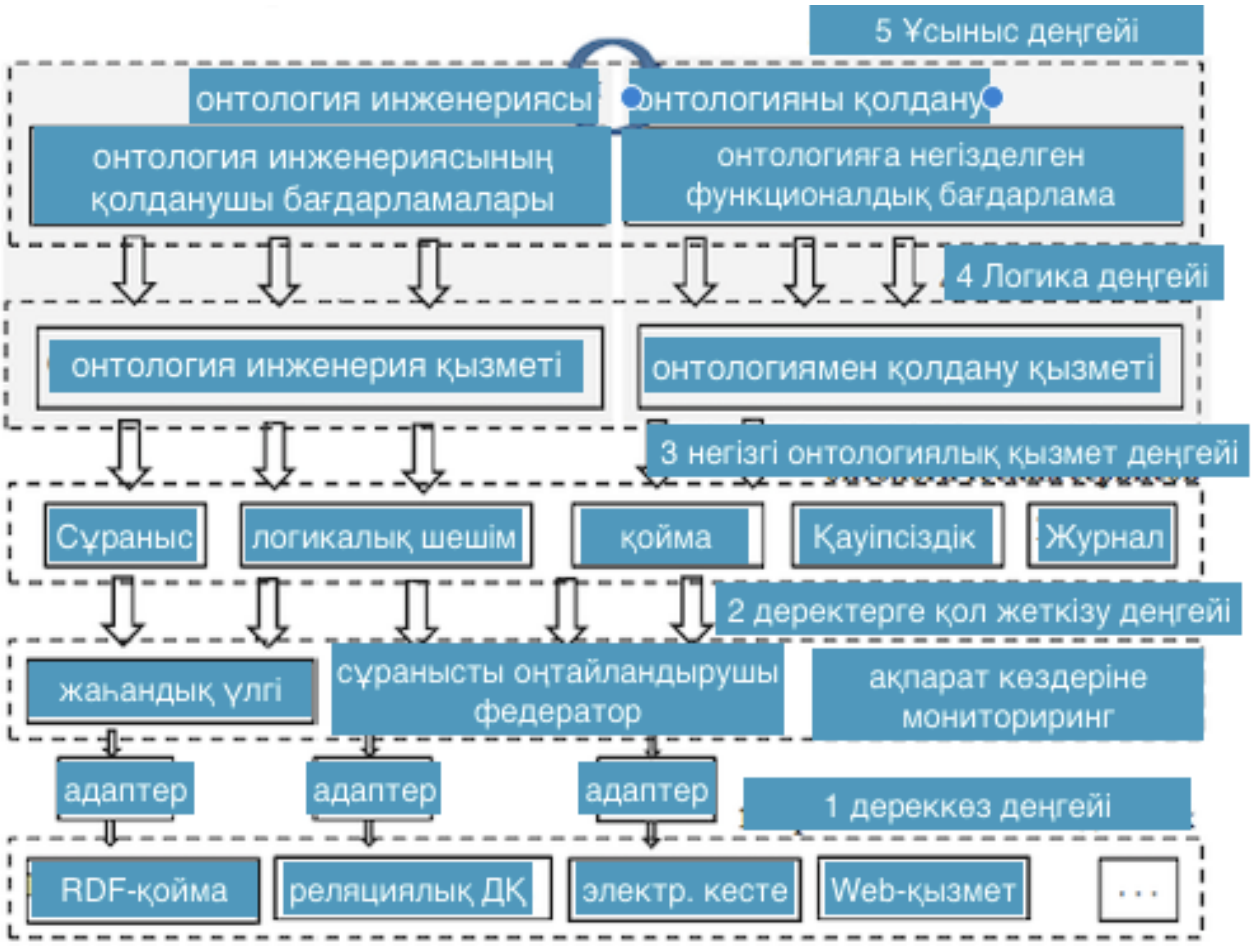
Семантикалық ДҚ мен технологиялардың негізінде ақпараттық жүйелердің жаңа түрлері дамытылып келеді (сурет 2.7). Архитектурада келесі деңгейлерді қамтиды:

1) әртүрлі дерек көздерден жиналған гетерогенді дерек қорлар.

2) гетерогенді деректермен ары қарай жұмыс істеу мақсатында тазарту, форматтау үрдістерінен өткізу.

Бұл деңгейдің объектілі үлгісі семантикалық деректерді таратушы (онтологиялық дерек беруші) немесе қарапайым ақпарат таратушы (көбінде бұл реляциялық ДҚ) сияқты ақпарат таратушыдан алынуы мүмкін.

Берілген деңгей API мен арнайы адаптерлардан тұрады. Деректерді таратушы адаптердің шынайы уақытта RDF-үлгісінде қолданылатын форматтардан деректердің өңделуін іске асыруға мүмкіндік беретін жүйе.



Сурет 2.7 – Семантикалық деректер негізіндегі ақпараттық жүйенің жалпы құрамы

3) базалық онтологиялық қызметтердің деңгейі келесідей құрамнан тұрады:

- Онтологияны тіркеу қызметі (жариялыным, ізденіс);
- Онтологиямен жұмыс жасайтын қызмет, мысалы элементтерді алу мен оларды өзгерту;
- Деректерді алу үшін жіберілетін сұраныс қызметі;
- Логикалық қортынды қызметі;
- Онтология мен жұмыс журналын жүргізу қызметі.
- Семантикалық ДҚ қауіпсіздігін қолдау қызметі. Семантикалық ДҚ қауіпсіздігін қолдау қызметі семантикалық ақпараттық жүйелердің жұмыс істеу үрдісінде аса маңызды рөл атқарады. Олар мынандай мүмкіндіктерге ие болуы керек:

- Қоймада сақталынған деректерге қолданушылардың қол жетімділігін қамтамасыз ету;
- ДҚ сұранысын орындауды бақылау;
- Жіберілген сұраныстарды орындау кезінде алынған логикалық қорытындыларды қамтамасыз ету.

4) Логика деңгейі онтология инженерия қызметі мен базалық операцияларды орындау қызметінен, онтология мен мета деректерден тұрады (түсініктер экземплярлары). Олар қолданылатын нұсқалар үшін өте тиімді болып келеді (өңдеу логикасын қосады) және нақтылы объектілі үлгілермен жұмыс істейді (нақтылы деректермен). Берілген қызметтер таңдауды орындау мен семантикалық деректерді басқару үшін отнологияның өмір сүру циклын туғызады (ontology lifecycle services).

Берілген деңгейдегі объектілі үлгілер әртүрлі таратушы жерлерден келетін деректерден тұруы мүмкін, мысалы қарапайым дерек таратушы (мысалы реляционды ДҚ), семантикалық деректерді таратушы (онтологиялық таратушы).

5) таныстыру деңгейі қолданушының графиктік интерфейсінен тұратын компоненттерден тұрады, мысалы web-қосымша мен desktop-қосымша, бұлар логика деңгейіндегі компоненттермен өзара әрекет етеді.

Қызметке деген сұраныс осы деңгейдегі қосымшамен жұмыс істеу барысында логика деңгейінде орындалады. Берілген қосымшалар екі топқа бөлінеді:

- онтологияның редакторларынан, браузерынан, онтологияның кітапханасынан тұратын инженериясын қолдаушы қосымша;
- порталдардан, электронды кітапханалардан, арнайы ақпараттық жүйелерден, шешімді басқарушыларды қабылдаушы жүйелерден, білімді басқару жүйесінен тұратын онтологияны қолданатын қосымшалар.

2.3.1 Білімге қатысты онтологиялық әрекет

Базалық ДЛ ALC [29] негізгі түсінігінің анықтамасын келтіреміз. Онтологияны сипаттайтын тілдер, мысалы OWL Lite, OWL DL және OWL2, күрделі ДЛ-да болады. Мысалы, OWL Lite тілінің математикалық негізі ДЛ SHIF(D), ал OWL2 тілінде – логика SROIQ(D). Бұл логикалар ALC базалық логикасының кеңейтілуі болып табылады. Көрсету жолының қысқа болуы мен жалпы жағдайы [30] терминологиясының көмегімен ALC логикасы көрсетілген.

Анықтама 2.1 N_C – түсініктер атауларының жиыны, N_R – қатынас атауларының жиыны. ALC-түсінігінің жиыны мыналар:

1. $\top$ (көп салалы түсінік), $\bot$ (бос түсінік) және $A \in N_C$ түсініктерінің барлық атаулары ALC –түсінігі болып табылады;

2. егер C мен D – ALC-түсініктері мен $r \in N_R$, онда $C \sqcap D$, $C \sqcup D$, $\neg C$, $\forall r.C$, $\exists r.C$ $\exists$ түсінігі ALC –түсінігі болып табылады.

Семантика ДЛ төмендегідей жолмен ALC логикасы үшін анықталатын интерпретацияның көмегімен беріледі.

Анықтама 2.2 Интерпретация деп $I = (\Delta^I, \cdot^I)$ жұбы аталады, ол Δ^I кішкентай жиынындағы әрбір ALC-түсінікті бейнелейтін $\cdot^I$ функциясы мен интерпретация

домені деп аталатын, бос болмайтын Δ^I жиынынан тұрады, ал N_R қатынасының әрбір атауы – $\Delta^I \times \Delta^I$ туындысының декарттік жиыны мынандай, кез-келген ALC-түсінігі үшін C, D мен r атауының қатынасы дұрыс:

$$\top^I = \Delta^I,\ \bot^I = \emptyset,$$
$$(\mathrm{C} \sqcap \mathrm{D})^I = C^I \cap D^I,\ (\mathrm{C} \sqcup \mathrm{D})^I = C^I \cup D^I,\ \neg \mathrm{C}^I = \Delta^I \backslash C^I,$$
$$(\exists \mathrm{r}.\mathrm{C})^I = \{x \in \Delta^I | \exists y \in \Delta^I : (x, y) \in r^I \wedge y \in C^I\},$$
$$(\forall \mathrm{r}.\mathrm{C})^I = \{x \in \Delta^I | \exists y \in \Delta^I : \text{егер } (x, y) \in r^I, \text{то } y \in C^I\},$$

Анықтама 2.3 түсінікті білдіретін аксиома деп $C \sqsubseteq D$ түріндегі бекітілуін атайды, мұндағы C мен D –ALC-түсінігі.

Анықтама 2.4 түсінікті енгізу аксиомасының жиын сапасы TBox, немесе онтологияның терминологиялық бөлігі деп аталады.

Анықтама 2.5 I интерпратациясы $C^I \subseteq D^I$ болғандағы $C \sqsubseteq D$ түсінігін енгізу аксиомасының үлгісі деп атайды.

Анықтама 2.6 I интерпратациясы TBox I үлгісі деп аталады, егер ол I түсінігінің барлық аксиомаларының үлгісі болса.

Анықтама 2.7 N_x – экземпляр атауларының жиыны болсын. Сонда экземплярларды бекітуді $x: C$ мен $(x, y): r$ деп белгілейді, мұндағы C – дұрыс ALC-түсінігі, r – байланыс атауы, ал $N_R x, y \in N_x$

Анықтама 2.8 экземплярлар туралы бекітілген қорытынды жиын ABox, немесе онтологияның фактологиялық бөлігі деп аталады.

Анықтама 2.9 I интерпратация $x^I \in C^I$ болған кездегі $x: C$ түріндегі экземпляр туралы бекітілген үлгі деп аталады. I интерпратациясы $(x^I, y^I) \in r^I$ болғанда $(x, y): r$ түріндегі экземпляр туралы бекітілген үлгі деп аталады.

Анықтама 2.10 I интерпратациясы A экземплярлары жайлы барлық ережелердің үлгісі болып табылған жағдайда ABox A үлгісі деп аталады.

Анықтама 2.11 I Tbox және A ABox жұптары базалық білім (I, A) (онтология) деп аталады.

Анықтама 2.12 I үлгісі мен A үлгісі бірдей уақытта I интерпратациясы болса K = (I, A) білім базасының үлгісі деп аталады.

Бағдарламалық жасақтамада онтологиялық шешімді пайдаданудың артықшылықтары:

- Салыстырмалы түрде жеңіл өзгерту, кеңейтуде ыңғайлы деректер үлгісі болып табылады. Онтологияның бұл қасиеті деректердің үлгілерін түрлендіруі мен тез уақытта кеңейтілуін қамтамасыз етеді.
- Бар онтологияның қайтадан қолданылу мүмкіндігі. Бұл артықшылығы деректердің сызбасын қайтадан жасамай тәжірибелік апробацияны өткен дайын шешімдерді қолдануға мүмкіндік береді.

2.3.2 Ғылыми ақпараттарды есепке алу, талдау жүйесінің үлгісі мен архитектурасы

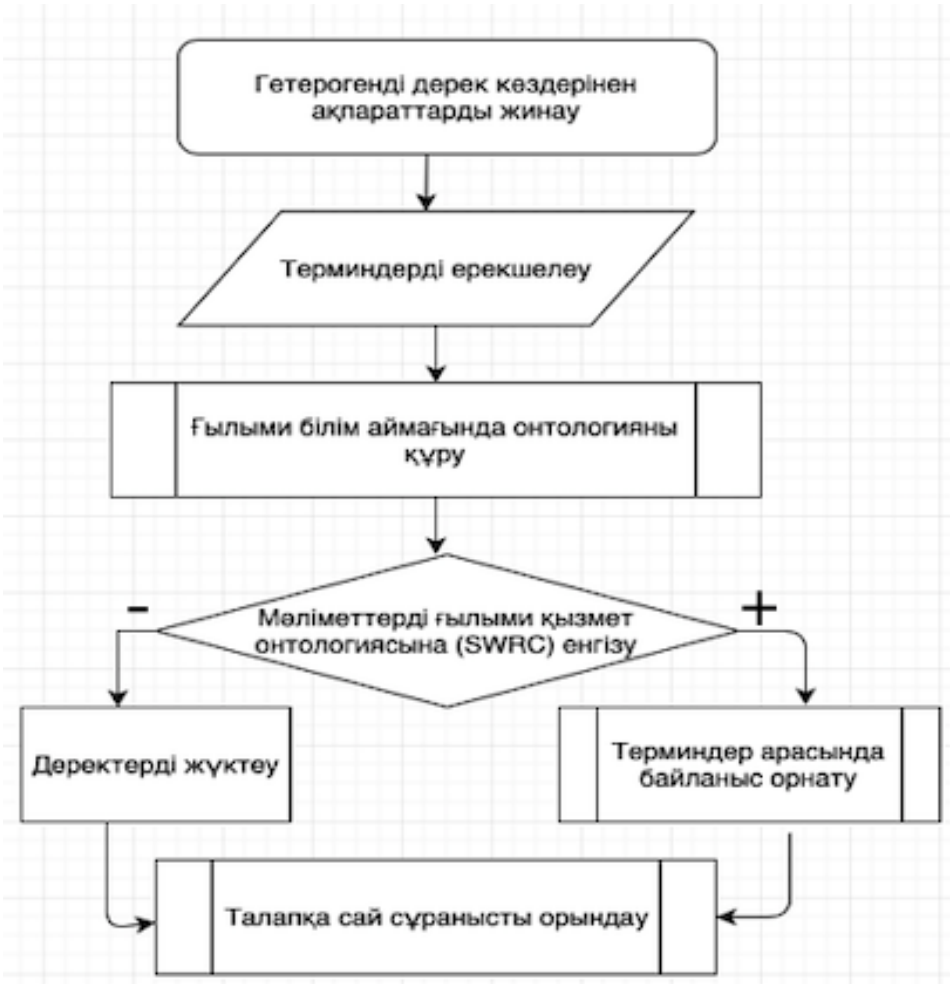
Автоматтандырылған жүйенің негізгі прототипі болып табылатын ғылыми-техникалық ақпараттың қорытындысы мен күрделі ұйымдастырылған есептеу үрдісінің жалпы формальды үлгісін қарастырамыз.

D ғылыми білім аймағы берілді деп есептейік (мысалы, Computer Sciense). I – білімнің осы аймағының шеңберіндегі ғылыми-техникалық ақпараттың бірліктерін сипаттау жиыны болсын. I жиынына ғалымның баспаға шыққан еңбектері кіреді.

Жүйенің негізгі мақсаты – іздеу-аналитикалық сұранысын орындау. Типтік сұраныстардың жиынын Q белгісімен белгілейміз. Тапсырма $I_q \subseteq I$ ғылыми-техникалық ақпарат бірлігінің сипаттамаларымен $q \in Q \quad r1: Q \to 2^I$ көрінісі арқылы беріледі.

Жүйенің жалпы сызбасы сурет 2.8 көрсетілген және олар мынандай үлгілерден тұрады:

- Білімнің осы аймағына бағытталған ғылыми-техникалық конференцияны мәтіндік сипаттаудан D ғылыми білім аймағын сипаттайтын терминдерді ерекшелеп алу;
- D ғылыми білім аймағындағы қарастырылатын онтологияны құру;
- Ғалымдардың мақалаларын деректерді жүктеу;
- Білім аймағында құрастырылған экземплярлар мен ғылыми салалардың нәтижелері жайлы жүктелген ақпараттардың арасындағы байланысты орнату;
- Алынған ақпаратқа аналитикалық сұранысты орындау;
- Жалпы сызба келесі сатылардан тұрады:

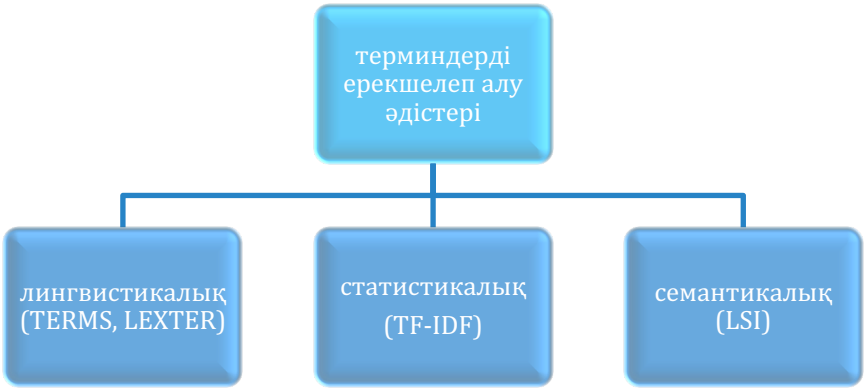


Сурет 2.8 – Ғылыми ақпаратты басқару жүйесінің жалпы архитектурасы

- D ғылыми білім аймағын сипаттайтын терминдерді ерекшелеп алу (кілт сөз);
- D ғылыми білім аймағының онтологиясын құрастыру;
- Ақпараттардан алынатын ғылыми саласындағы деректерді жүктеу;
- Құрастырылған онтология түсінігі мен қолданушылардың ғылыми қорытындыларының арасындағы байланысты орнату;
- Сұраныстардарды орындайтын ақпараттың қорытындысын қамтамасыз ету.

2.3.3 Білім аймағын сипаттайтын терминдерді ерекшелеу

Мәтіндік құжаттардан ақпаратты алу ақпараттың көлемін өсуімен тығыз байланысты. Термин деп құжаттың қысқа сипаттамасын білдіретін не оның ақпаратын классификациялау мен класстеризациялау сияқты өңдеуде қолданылатын сөздер мен сөз тіркестерін атайды. Көптеген зерттеушілердің назары терминдерді ерекшелеп алу тапсырмасында. Сурет 2.9 бір ғана сөзден тұратын сөздерді ерекшелейтін алгоритмдердің тізімі [31] берілген.



Сурет 2.9 – Терминдерді ерекшелеу

Лингвистикалық жол мәтіннің семантикасы мен морфологиясының негізінде сөз тіркестері мен кілт сөздерін ерекшелеуге көмектеседі (мысалы, WordNet базасы). Лингвистикалық өңдеудің нәтижесі - тақырыптар аймағының мағыналы және қалыпқа келтірілген сөздерінен тұратын құжаттар жиыны болып табылады [32].

Статистикалық жол құжатқа енетін ақпарат сөздердің маңыздылығын арттырады, яғни барлық сөздерге баға беруге әсерін тигізеді. Статистикалық жолда үлгінің дұрыстығына негізделген тәсілдерді кеңінен қолданылады [33,34]. Статистикалық жолмен әрекет ету арқылы мықты математикалық аппаратарды қолдану арқылы алгоритмдердің әсерлі екендігін теориялық тұрғыдан дәлелдеуге болады, бірақ олар сөздердің сөйлемде пайда болу тәуелдігіне сүйенеді.

Семантикалық критерийді адамдар ғана қанағаттандыра алады. Статистикалық әдістерді қолдану арнайы мәтіндердегі терминдерді объективті және дәлірек таңдауға мүмкіндік береді [34,p. 302].

Анықтама 2.13 берілген диссертацияда термин деп - оның бір немесе бірнеше тақырыптарына сәйкес келетін құжатты сипаттайтын жұп сөздерді

атаймыз. Тапсырманың формальды шешімі келесі жолмен іске асады. $r_1, \dots r_n$ айдарларына бөлінген Doc көптеген құжаттары берілсін делік. Тапсырма $Terms$ терминдерінің жиынынан тұрады. Әрбір $A \in Terms$ термин реттелінген екі сөзден тұрады: $A = (A_1, A_2)$ Терминдерді ерекшелеу алгоритміне қойылатын талаптар:

- Ғылыми білім аймағындағы бағытты сипаттайтын терминдерді құжаттан шығарылуы тиіс.
- Шығарылатын терминдердің нақтылығы терминдерді ерекшелеу алгоритмінің апробациясынан өтпеу керек.
- Алгоритм білім алуда өңдеуге көп күш жұмсамай ашық дерек көзден қажетті білімді ала алатын мүмкіндігі болу керек.
- Алгоритм берілген пәндік аймаққа орнату бойынша эксперттердің қол еңбегінің көп күшін қажет етпеуі керек.
- Алгоритм үшін деректердің алынатын жері ашық және жаңартылып отырылуы керек.
- Алгоритм үлгілі архитектурадан тұруы керек.
- Білімнің нақты бір аймағына алгоритмнің автоматты орналасу мүмкіндігі болуы керек.
- Берілген диссертацияда ғылыми ақпараттың ашық дерек көзі ретінде жаңа конференциялардың жариялануы қаматамасыз ететін құжат Call for Papers (CFP) (сурет 2.10) қолданылады.

Computing Conference 2022 - Call for Papers
14-15 July 2022

We'd like to invite you to submit your papers for the Computing Conference 2022 to be held from 14-15 July 2022 in London, UK.

We are aware that the situation regarding COVID-19 is a cause for apprehension. There will be an option for virtual participation (with reduced registration fee), for anyone who cannot or chooses not to travel.

Computing Conference is a research conference held in London since 2013. The conference series has featured keynote talks, special sessions, poster presentation, tutorials, workshops, and contributed papers each year. The goal of the conference is to be a premier venue for researchers and industry practitioners to share new ideas, research results and development experiences in computing and related fields.

Conference Tracks include Computing, Artificial Intelligence, Machine Vision, Security, Communication, Ambient Intelligence, e-Learning & Technology Trends.

Important Dates:
Paper Submission Due: 01 October 2021
Acceptance Notification : 01 November 2021
Camera Ready Submission : 15 January 2022
Conference Dates : 14-15 July 2022

Complete details are available on the conference website : http://saiconference.com/Computing

Computing Conference 2022 proceedings will be published in Springer series "Lecture Notes in Networks and Systems" and submitted for consideration to Scopus, Web of Science, DBLP, INSPEC, WTI Frankfurt eG, zbMATH, SCImago

Please get in touch if you would like to contribute in more ways; and feel free to circulate this information among your colleagues/students.

Looking forward to your contributions.

Regards,
Kohei Arai
Program Chair, Computing Conference 2022

Сурет 2.10 – Ғылыми конференцияның жариялану анонсы

Электронды почталық сілтемелердің тізімі, әдетте, ғылыми білімнің белгілі бір бөліміне арналады. Бұл факт мынаны білдіреді, CFP құжатының қажетті айдары ғылымның әртүрлі аймағына арналған сілтеменің бірнеше тізімі жайлы деректер ретінде қабылдау жолы арқылы алынуы мүмкін. Осындай жолмен конференциялардың жариялануын пән аймағы жайлы ақпарат таратушы ретінде қолдану ғылыми білім аймағындағы берілген шеңбердегі құжаттарды таңдауға үйретететін дайындық бойынша қол еңбегінің экспертінің минимизациясын қамтамасыз етеді.

2.4 Ғылыми білім аймағындағы онтологияны құру

Ғалымдардың ғылыми еңбегінің нәтижелері жайлы деректерге күрделі аналитикалық сұранысты орындау мүмкіндігін қамтамасыз ету үшін білімнің осы аймағы жайлы түсінік пен ғылыми ақпараттың бірлігін (мысалы, мақала) байланыстыратын және білім аймағын сипаттайтын қосымша ақпарат қажет.

Мәтіндер коллекциясының негізінде онтологияның құрылуына арналған формальды тапсырма келесі жолмен анықталады.

Анықтама 2.14 $O = (I, A)$ онтологиян құрауға арналған тапсырма оның формальды (автоматизацияланған) қорытындысының негізінде қалыптасуын және пән аймағының онтологиясын жасаушыны қызықтыруға тұрарлықтай Doc мәтіндер коллекциясының экспертін (немесе эксперт тобы) таңдаудан тұрады: N_C түсінігінің атаулар жиыны; N_R қатынасының атаулар жиыны; N_x экземплярының атаулар жиыны; түсініктерді енгізу аксиомасының $I = TBox$ қорытынды жиыны (онтологияның терминологиялық бөлімі); экземплярларды бекітудің $A = ABox$ қорытынды жиыны (онтологияның фактологиялық бөлімі).

Келесі жолмен анықталатын онтологияны толтыру тапсырмасымен ұқсас болып табылады.

Анықтама 2.15 N_C түсінік атауларының жиыны, N_R қатынасының атауларының жиыны мен түсініктерді енгізу аксиомасының $I = TBox$ қорытынды жиыны берілді деп алайық. Онда $O = (I, A)$ онтологиясының толықтырылу тапсырмасы экземплярларда бекітілген $A = ABox$ қорытынды жиыны мен N_x экземплярлар атауларының жиынын ақпарат негізінде құрау мен Doc мәтіндерінің коллекциясын таңдаудан тұрады.

Онтологияны құрастыруда түсініктердің экземплярлары мен олардың арасындағы байланыс орнатылады. Онтологияның терминологиялық бөлімін құрастырудың қорытындысы тапсырманы орындау үшін арналған әдіс ретінде термин мен олардың байланысы, зерттеу аймағы ретінде осындай түсініктердің жиыны, әдістер, тапсырмалар, шешімдер, терминдер болуы мүмкін, зерттеу аймағы арнайы тапсырмалардан тұрады, термин әдістемеде қолданылады, термин зерттеу аймағындағы кілттік түсінік болып табылады. Онтологияны құру мен толтыру келесі себептер бойынша маңызды.

- Онтологияны құрастыру мен толтыру табиғи тілдегі мәтіндерден ақпараттарды автоматты түрде ерекшелеп алу алгоритімін жасауды қажет етеді. Корпорацияның 80% деректері табиғи тілдегі мәтінде орналасқан[35]. Сондықтан интеллектуалды алгоритмдер осы үрдісті автоматтандыруда маңызды.

- Толтырылған онтология семантикалық тор үшін дайын ақпараттық ресурс болып табылады (Semantic Web [35,p. 430]). Ол үшін мета деректердің генерациялық автоматты қажет. Онтологияны толтыру семантикалық аннотацияның құрастырылуына септігін тигізеді, яғни, веб-парақшадағы ақпаратты онтологияға айналдырады.

Ғылыми ақпаратты есепке алып, қорытынды жасау үшін D ғылым аймағындағы білім бойынша берілген $O_D = (I_D, A_D)$ онтологиясын құру қажет. 2.14 анықтамасына сәйкес Doc мәтіндер коллекциясын таңдап, содан кейін N_C^D түсінігінің атауларының жиыны, N_R^D қатынасының атаулары, $TBox_D$ аксиомасының N_x^D экземплярлар жиыны мен $ABox_D$ бекітілуі жасалынған жүйеге қатысты жалпы талаптардан шыға отырып, осы онтология мен оны құрастыру алгоритміне қойылатын талаптарды тізбектейміз.

- Алгоритм терминдердің арасынан иерархиялық және ассоциативті байланыстың болу керек. Онтологияда білім аймағындағы айтарлықтай көп ақпарат болуы керек.
- Онтология ғылыми білімнің берілген аймағындағы маңызды жағдайды көрсетуі керек.
- Алгоритм білім алуды қажет етпеуі керек, не болмаса оны өңдеуге көп күш жұмсалынбайтын ашық дерек көздерінен қажетті білім алу жолдарының болуы керек.
- Алгоритм берілген пәндік аймақта эксперттің қол еңбегін қажет етпеуі керек.
- Алгоритм үшін деректер көзі қол жетімді болып, уақытылы жаңартылып отыруы керек.
- Алгоритмнің білімнің нақтылы бір ауданына қатысты автоматты түрде орнатылу мүмкіндігі болуы керек.

Білімдер доменінің онтологиясының қарапайым мысалдары ретінде Әмбебап ондық жүйелеу (ӘОЖ-УДК) және Есептеу техникасы қауымдастығы (ЕТҚ-CCS - Computing Classification System) сияқты классификаторлар табылады. Диссертация тапсырмасын шешуде қолданыстағы классификаторларды қолдану жоғарыда аталған талаптарға сәйкес келмейді.

Ол жерде концепциялар арасындағы иерархиялық қатынастар ғана орындалады. Ал аналитикалық сұраныс ассоциациялық қатынас арқылы орындалады және жіктеуіштер салыстырмалы түрде сирек жаңартылады.

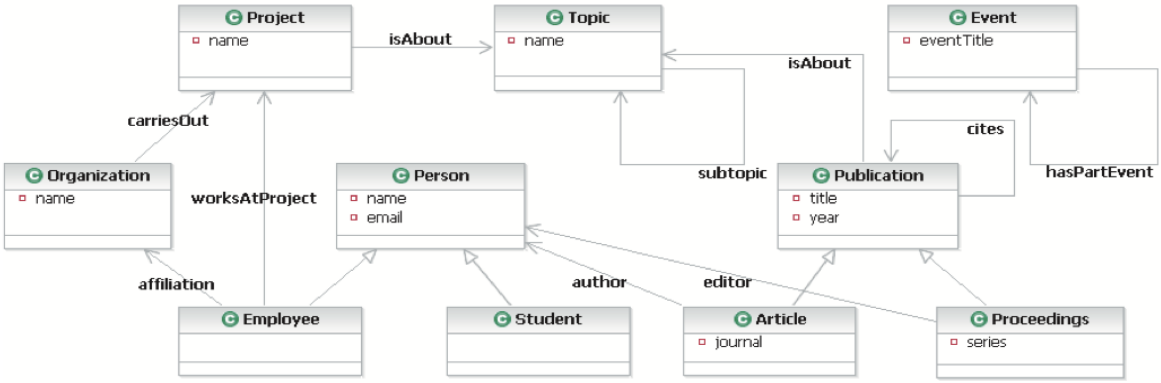
Ғылыми білім аймағындағы онтологияны құрастыру үшін жеті деңгейден тұратын ғылыми онтология жасаушы алгоритмі қолданылады:

- Әр түрлі білім салаларына ортақ ұғымдардан тұратын N_D^C тұжырымдамаларының атауларының жиынтығын құру;
- Терминдерді бөлу алгоритмін қолдана отырып, D ғылыми білімінің белгілі бір саласын сипаттайтын терминдерді бөліп көрсету;
- Терминдерді фильтрациядан өткізу;
- Терминдердің арасындағы ассоциативті байланыстарды ерекшелеу;
- Терминдер иерархиясын құрастыру;
- Терминдерді қазақ/орыс тіліне аудару;
- Онтологиялық түсінік бойынша терминдерді саралау.

2.5 Деректерді жүйеге жүктеу

Ақпараттарды жүйеге жүктеу үрдісі қолданушының енгізген бірінші өңдеу жұмысының жүйесі мен оның ішінен қажетті ақпаратты таңдап алудан тұрады, мысалы, қызметкер қатысқан конференциялар мен жобалар жайлы деректер, мақалалардың атауы, орыны, күнінің белгіленуімен қызметкердің жарияланған мақалалар тізімі. Осы деректерді көрсету үшін берілген диссертацияда SWRC онтологиясы қолданылды.

Ғылыми саланың жалпы сипаттамалық бағытымен (берілген мақала, конференцияларға жасалатын қорытындысыз) қолданылатын құжаттарда болатын деректерді формалды түрде сипаттау үшін қажет көптеген түсініктер осы онтологияда қалыптасқан. SWRC онтологиясының кілттік түсінігі сурет 2.11 берілген. SWRC онтологиясында сөздіктердің осы түрі бар [36]. Мысалы, Dublin Core онтологиясы қарастырылып отырған тапсырманы шешу үшін қолданылуы мүмкін.



Сурет 2.11 – SWRC онтологиясы (фрагмент)

Ғылыми саланың $O_S = (T_S, A_S)$ онтологиясын қарастырамыз. Бұл онтология SWRC онтологиясына сәйкес келеді. N_C^S , N_R^S мен N_x^S – оның түсініктерінің атаулар жиыны, соған сәйкес байланыс пен экземпляры. Ғылыми-техникалық ақпараттың бірлік сипаттамасының мәтіндік I жиыны берілсін делік. Қарастырылатын деңгейдің мақсаты I алынған ақпараттардың негізіндегі байланыс мәні мен түсінігінің экземплярлары O_S онтологиясын толықтыру болып табылады. 2.15 анықтамасына сәйкес бұл тапсырма келесі жолмен құрастырылады.

Егер $p \in I$ – мақала болса, онда O_S онтологиясында Person, Journal, Proceedings және creator, title және year байланыстары болсын. Онда p элементі үшін тізбектелген түсініктердің экземплярларын жасау қажет.

Осындай жолмен деректердің жүктелуі қызметкерлер енгізетін ақпараттардан ерекшеленіп алынатын экземплярлармен O_S онтологияның толықтырылуынан тұрады. Деректерді енгізудің келесі тәсілдері жоспарланады:

- библиографиялық сілтемелерді реттеу;
- жүктелетін мета деректерді реттеу (BibTeX, MathML, LaTeX, FinXML);
- жолдарды қолмен толтыру.

Деректерді жүйеге енгізу үш сатыдан тұрады.

1 қадам. Қолданушы алынатын деректерді енгізеді, мысалы библиографиялық сілтемелер, BibTeX-жазбалар. Жүйе автоматты түрде қолданылатын деректерден ақпаратты алып тастайды, және қолданушыға толтыратын жолын көрсетеді, ол автор, атауы, жарияланған жері, басқа да арнайы жолдар жұмысты сипаттайды.

2 қадам. Қолданушы жолдарға жазған мәтінді қайта түзете алады.

3 қадам. Қойма жүйесінде орналасқан объектілермен бірге жолдардың мәнінде жазылған сөз сәйкес болуы керек. Бұл саты деректерді енгізу барысында рұқсат алуды мақсат етеді.

Сәйкестік деңгейі аса маңызды болып табылатынын атап өту керек, себебі:

- қосымша ақпараттарды бөлек іздеуді іске асыру қажеттілігіне кедергі келтірмейтін жағдайдағы, қолданушының көмегімен ғалымдардың аты-жөні әртүрлі екендігі жайлы мәселені дұрыс шешуді қамтамасыз етеді;

- енгізілген мәтінді түзету мүмкіндігі міндетті түрде болу керек.

Төменде бірінші сатының толығымен сипаттамасы көрсетілген.

Библиографиялық сілтемелерді реттеу. Библиографиялық сілтемелерден ақпаратты алу ақпаратты құрамы жасалынбаған мәтіннен алу тапсырма болып табылады [37]. Берілген библиографиялық сілтемелердің реттелу әдісін тестілеудің нәтиже көрсеткіші бойынша аса жоғарғы әсерін көрсеткен Conditional Random Fields (CRF) [38] алгоритмі. FreeCite бағдарламалық кешенінде CRF алгоритмі тегін таратылатын CRF++ кітапханасы көмегімен жүзеге асырылады. Ол келесі деңгейлерден тұрады.

- кіру нүктесінде библиографиялық сілтеме бір не бірнеше жол ретінде беріледі. Мысал ретінде мына сілтемелерді келтіруге болады:

```
R.Uskenbayeva, T.Chinibayeva. Model, data integration algorithms of information systems based
on ontology // Journal of Theoretical and Applied Information Technology E-ISSN 1817-3195
ISSN 1992-8645 Vol.99 May 2021 No 09. pp 2125-2143
```

- Жеке сілтемелерді ерекшелеу. Егер кіріске жаңа жол таңбасымен бөлінген бірнеше библиографиялық сілтемелер берілсе, олар бөлінеді. Осыдан кейін алгоритм сілтемелердің әрқайсысына дербес қолданылады.

- Сілтеме токенизациясы. Бұл деңгейде мәтіндік сілтеме реттілік атомдық токен-элементтеріне бөлінеді. Келтірілген мысалда сілтеме мынандай болып өзгереді:

```
“R.Uskenbayeva, ”, “T.Chinibayeva”, “Model, ”, “data ”, “integration ”, “algorithms ”, “of ”,
“information ”, “systems ”, “based ”, “on ”, “ontology ”, “Journal ”, “of ”, “Theoretical ”,
“and ”, “Applied ”, “Information  ”, “Technology ”,  “E-ISSN ”, “1817-3195”, “ISSN ”,“1992-8645 ”,
“Vol.99 ”, “Vol.99 ”, “2021 ”, “No 09. ”, “2125-2143”.
```

- Әрбір токеннің қасиетіне сипаттама беру. Бұл деңгейде әрбір токенді токен префиксы (алғашқы бірнеше символдар), суффиксы, алдын-ала құрастырылған сөздікке жатуы (мысалы, дат және географиялық атаулары бар сөздіктерге), және басқа да сипаттамаларға байланысты қасиеттері анықталады. Келтірілген мысалды сілтеме келесі түрде беріледі:

```
[[“T.Chinibayeva”, “,”, “T.”, “T.Ch”, “Chin”, “Chinib”, “,”, “bay,”, “yev,”, “yeva,”,
“nibayeva”, “InitCap”, “nonNum”, 0, “noMaleName”, “noFemaleName”, “noLastName”,
“noMonthName”, “noPlaceName”, “noPublisherName”, “noEditors”, 0, “contPunct”, “notInBook”],
[“I.S.,”, “,”, “I”, “I.”, “I.S”, “I.S.”, “,”, “.,”, “S.,”, “.S.,”, “is”, “AllCap”, “nonNum”,
16, “noMaleName”, “noFemaleName”, “noLastName”, “noMonthName”, “placeName”,
“noPublisherName”,“noEditors”, 0, “contPunct”, “notInBook”] .....].
```

- CRF алгоритмін қолдану арқылы токенді классификацияға жіктеу. Бұл деңгейде әрбір токенге сілтеменің жазылуын білдіретін белгілердің бірі қойылады, мысалы, автор, атауы, журналдың аты, парағы. Келтірілген мысалда мынандай белгілер болады:

```
“author”, “author”, “author”, “author”, “author”, “author”, “author”, “author”,
“author”, “author”, “author”, “author”, “author”, “title”, “title”, “title”,
“title”, “title”, “title”, “title”, “title”, “title”, “title”, “title”, “title”,
“title”, “journal”, “journal”, “journal”, “journal”, “journal”, “journal”, “date”,
“pages”.
```

- Жолдардың мәнін стандартты түрге келтіру. Бұл деңгейде оларды стандартты унифицирленген түрге келтіру үшін жолдардан алынған мәнді қосымша өңдеу жұмысы жүргізіледі. Келтірілген мысалда өңделу жолы мынандай болады:

```
[“R.Uskenbayeva, ”,  “T.Chinibayeva ”] => “R.Uskenbayeva, T.Chinibayeva”;
“09: 2125-2143” => {:volume =>09, spage => 2125, :epage => 2143}.
```

BibTeX-жазбаларын реттеу. Жүйеге ақпаратты енгізудің мүмкін болатын келесі тәсілі BibTeX форматындағы жазбаны енгізу. Библиографиялық сілтемені енгізуге сәйкес жүйе автоматты түрде жолдарға енгізілген жазбаларды реттейді. BibTeX құрамдық формат болғандықтан бұл үрдіс интеллектуалды алгоритмдерді қолдануды қажет етпейді. Көмекші құрал ретінде Python тілінде жазылған pybtex кітапханасы қолданылады.

Деректерді қолмен енгізу. Жоғарыда сипатталған жүйелерге деректерді енгізудің автоматтандырылған әдісі мақалалар жайлы ақпараттарды енгізу барысында ғана мүмкін болмақ. Конференциялардағы есеп, патенттер, журналдың редколлегиясына қатысуда, есеп берулер мен басқа да ғылыми салалар жайлы деректерді енгізу кезінде қолданушыға атауы, автордың есімі, және соған ұқсас жолдарды қолмен толтыру қажет.

Деректерді библиографиялық сілтемені BibTeX-жазбалары ретінде енгізу барысында қолданушы бірнеше мақалалар жайлы ақпараттың формасына енгізуіне болатынын ескере кеткен жөн. Алынған барлық деректерді өңдеу үшін кезегімен орналастырылады және қолданушы кезегі бойынша бірінші жұмысты қосатын екінші қадамға көшеді. Қосып отыру кезегі жүйемен істелетін жұмыстардың арасында сақталады және қолданушы кезегі бойынша кез-келген уақытта келесі мақаланы қосу үшін қайтадан оралуына болады.

2.6 Ғылыми білім аймағындағы онтология мен жүктелген деректердің арасындағы байланысты орнату

Білім аймағындағы қалыптасқан үлгі мен ғалымдардың ғылыми еңбектерінің нәтижесінің арасындағы байланыс аналитикалық сұраныс көмегімен жүзеге асады.

Білімді көрсету үлгісінің қолданылу шеңберінде ғылыми білім аймағындағы онтология O_D мен жүйеге келіп түсетін деректердің ішінен алынатын ақпараттардың негізінде экземплярлармен толықтырылған ғылыми білім аймағындағы онтология O_S байланысы онтологиялардың бірігуін болады. Ол онтологияның экземплярларының арасындағы байланыстың (сілтеме) қалыптасуына алып келеді. Бұл деңгейдің мақсаты экземплярлардың кішірек жиынындағы N_x^S (мақала, есеп тағы басқа еңбек нәтижелері) экземплярларына сәйкес келетін $SD: N_x^S \to 2^{N_x^D}$ көрінісін құрастыру болып табылады, яғни осы нәтижелерді сипаттайтын D ғылыми білім аймағында берілген терминдерді құрайды. Бұл көріністі N_x^D (мысалы, әдістердің, ғылыми бағыттардың, тапсырмалардың атаулары) жиынындағы білім аймағының терминдері мен N_x^S жиынынан (мысалы, мақала) алынған ғылыми ізденістің нәтижесін байланыстыратын O_D мен O_S онтологиялардың экземплярларының арасындағы swrc:isAbout байланыс түрінің жиыны ретінде қарастыруға болады. Бұндай байланыстың мысалы ретінде журналдағы басылымның O_S онтология экземпляры арасында swrc:isAbout байланыс болуы мүмкін.

Ғылыми жарияланымдар келесідей әдіспен жіктеледі:

- *Әрбір санатқа сәйкес жарияланатын мақала түседі.* Бұл әдістің кемшілігі, біріншіден, журналдардың өз тақырыптарында көрсетілгеннен гөрі кеңірек білім саласын қамтитындығында. Екіншіден, бұл жіктеу әдісі білім саласының тар бағыттарын ескермейді.
- *Бірлескен дәйексөз.* Егер екі мақалаға бірдей жарияланымға сәйкес келетін болса, онда бұл екі мақала ұқсас болып саналады. Бұл әдіс жаңа тенденциялар мен зерттеу бағыттарын анықтау үшін тиімді.
- *Бірлескен кері дәйексөз.* Бұл ретроспективті әдістің идеясы егер екі жарияланым да бірдей мақалаға бағытталса, онда олар ұқсас болып саналады.
- *Жарияланымдардың атауларында қолданылатын сөздердің жиілігін талдау.* Осы әдістердің шеңберінде әрбір мақала үшін сөздік портрет құрастырылады – ол бұны сипаттайтын сөздердің тізімінен тұрады [39].
- *Жарияланымдарға арналған веб-сілемелердің нәтижесі.* Бұл әдіс Интернеттегі мақалаларға бірлескен дәйексөздерді талдау идеясын бейімдеу болып табылады.
- *Машинамен оқу алгоритмі.* Бұндай алгоритмдер үшін жарияланымның атауы, авторлардың тізімі, жарияланған жері жазылған сөздер сипаттама болып табылады [40]. Кеңінен қолданылатын классификация әдісі тірек векторлық машина болып табылады [41]. Мақалалардың атауында көптеген әртүрлі сөздер кездесетінін ескере кетейік. Сөздердің мөлшерін азайту мақсатында зерттеу жұмысында ұқсас сөздерді класстерлерге біріктіруді ұсынады.

- *Жолдардың арасындағы қашықтыққа негізделген әдістер, мысалы* [42]. Бұл класстардың ішінен алынған ең қарапайым әдіс зерттелініп отырған диссертациялық жұмыста қолданылған.

Ғылыми мақалалардың іріктелуіне арналған көптеген зерттеулердің көбі негізгі ақпарат таратушы ретінде басқа жұмыстарға арналған сілтемелер қолданылады.

O_D (білім аймағындағы термин) онтологиясының $t \in N_x^D$ экземпляры мен O_S (мысалы, мақалалар) онтологиясының $e \in N_x^S$ экземплярының арасындағы Sim семантикалық жақындық деңгейін анықтау үшін берілген диссертацияда мынандай формула қолданылады:

$$Sim(e,t) = sim_{edit}(title(e),t),$$

мұндағы $title(e)$– мақаланың атауы, ал $sim_{edit}(s_1,s_2) = \frac{1}{1+editDist(s_1,s_2)} - s_1$ жолының s_2 жолына айналуы үшін қажет түзетулер санынан тең (қою, өшіру, ауыстыру) $editDist(s_1,s_2)$ Левенштейн қашықтығының негізінде [40,p. 156] жалғастырылатын жолдардың ұқсас функциясы. Егер $Sim(e,t)$ функциясының мәні C_{sim} константтың мәнінен асатын болса, онда онтология экземпляры мен ғылыми мақаланың арасында swrc:isAbout байланысы орнатылады. Формуланы қолдануға мысал ретінде «Жеке мәліметтерді басқару мәселесі» атты мақала мен «Жеке мәлімет» онтологиясының экземпляры арасындағы автоматты түрде орналасқан байланысты келтіруге болады. Бұл жолдардың арасындағы қашықтықтар 22 тең, олай болса $Sim(e,t)$ функциясы 0.04 тең.

Пәндік аймақтағы онтологияның экземпляры мен мақаланы сипаттаудың арасындағы семантикалық байланысты анықтау әдісі семантикалық байланыс көрініс табатын классификацияның алгоритмдерін қолдану жолы жақсаруы мүмкін [43], сонымен қатар қосымша ақпараттың қолданылуы да жақсартылады, мысалы конференция тақырыбы мен бірлескен авторлардың қызығушылықтарын айтуға болады.

2.7 Деректерге қатысты аналитикалық сұраныстарды орындау

Деректерге қатысты аналитикалық сұраныстарды орындау білім аймағын сипаттайтын үлгілерді іске асыратын бағдарламалармен жүйені қолданушылардың әрекет ету барысында қамтылады. Бұндай үлгі білім аймағы жайлы жалпы ереже сияқты, осы аймақтағы ұйымның қызметкерлерінің ғылыми зерттеу нәтижелері жайлы жалпы ақпараттың қосындысынан тұрады. Сол кезде сұраныстың автоматты түрде қайта жазылуын іске асыру қажеттілігі туындайды.

SPARQL тілінің синтаксисін көрсету мақсатында бағдарламалық қамтамасыз етуді (“Software Engineering”) жасауға арналған 2020 жылғы жарияланымдарды алуға мүмкіндік беретін сұранысқа мысал келтіреміз.

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX swrc:<http://nauka.iitu.kz/ontologies/swrc#>
PREFIX cs:<http://nauka.iitu.kz/ontologies/computer_science#>
PREFIX dc:<http://purl.org/dc/elements/1.1/>
PREFIX xsd:<http://www.w3.org/2001/XMLSchema#>
SELECT DISTINCT ?pub
WHERE {
?pub a swrc:Publication.
?pub swrc:year 2020.
?pub swrc:isAbout cs:Software_Engineering.
}
```

Бұл сұраныстарды құрастыру әртүрлі интерпретацияларға мүмкіндік береді. SPARQL формальды көрінісінің мүмкін болатын нұсқаларынан біреуін қарастыру үшін көмекші анықтама енгіземіз.

Анықтама 2.16 O_D онтологиясының екі экземпляры $x_1, x_2 \in N_x^D$ берілсін делік. x_2 экземпляр x_1 туындысы деп алайық, бұл экземплярлардың арасында $isSubTerm: x_2 isSubTerm\ x_3 \quad x_3 isSubTerm\ x_4$, … , $x_{m-1} isSubTerm\ x_m$, $x_m isSubTerm\ x_1$ иерархиялық байланыстың тізбегі бар $x_3, \dots x_m \in N_x^D$ экземплярының бос болмайтын жиыны бар, не болмаса x_2 $isSubTerm$ x_1бекітілген ережесі бар.

Жеке ғалымдардың ғылыми еңбегінің нәтижесі жайлы деректермен толтырылған O_S ғылыми еңбегінің онтологиясы мен D білім түрінің O_D аймағының онтологиясын қолдану кезінде SPARQL арқылы көреміз. Ары қарай берілген SPARQL тіліндегі барлық сұраныстарда есімдердің келесі кеңістігі бар:

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX swrc:<http://nauka.iitu.kz/ontologies/swrc#>
PREFIX cs:<http://nauka.iitu.kz/ontologies/computer_science#>
PREFIX dc:<http://purl.org/dc/elements/1.1/>
PREFIX xsd:<http://www.w3.org/2001/XMLSchema#>
```

swrc атаулар кеңістігі O_S ғылыми еңбегінің онтологиясының элементтерін көрсету үшін қолданылады (еске салайық, SWRC онтологияның кеңейтілген нұсқасы болып табылады), ал cs атауларының кеңістігі – пәндік аймақ сияқты информатика білімінің аймағындағы онтология (Computer Science). Dublin Core онтологиясын dc есімдерінің кеңістігі білдіреді. $T = \{t_1, \dots t_n\}$ терминдер жиыны D білім аймағында жеке ғылыми бағдарды анықтайды. ғылыми еңбектердің нәтижесінің барлығы SWRC қолданылатын онтологияға қосылатын Result классына жатады, кейбір сұраныстарды орындау үшін SPARQL қолданылады.

Білімнің қызықтыратын аймағындағы шеңберінде белсенді зерттелінетін бағыттардың тізімі. Бұл сұранысты былай деп жазайық: соңғы жылға (2020) ғылыми еңбектің сәйкес нәтижесі шығарылған терминдердің барлығын реттеп шығару. Оны SPARQL тілінде формаландырамыз. Алдымен соңғы жылда ғылыми еңбектің нәтижесіне сәйкес келетін мазмұнындағы барлық терминдердің (қайталануы да мүмкін) $Terms$ жиынын құрастырамыз.

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX swrc:<http://nauka.iitu.kz/ontologies/swrc#>

SELECT ?term
WHERE {
?term a cs:term.
?res a swrc:Result.
?res swrc:isAbout ?term.
?res swrc:year 2020.
}
```

Алынған Terms терминдердің жиынын әрбір жеке дара элементтің қайталану саны бойынша азаюына қарай реттеу қажет. Бағытты анықтайтын және реттелген тізімнің басында орналасқан терминдер білімнің қызықты зерттеу аймағында белсенді түрде қарастырылуда.

Жақын уақытта пайда болған бағыттардың, әдістердің, тәсілдердің тізімі. Бұл сұранысты былай деп жазайық: алдыңғы жылдың ешбір қорытындысымен сәйкес келмейтін, соңғы жылдың қорытындысымен ұқсас болып келетін барлық терминдерді көрсету керек; оларды терминдер санаты бойынша топқа бөлу керек. Оны SPARQL тілінде формаландырпмыз. Алдымен соңғы жылда ғылыми еңбектің нәтижесіне сәйкес келетін мазмұнындағы барлық терминдердің (қайталануы да мүмкін) $Terms$ жиынын құрастырамыз. Егер термин қандай да бір классқа жататын болса, онда ол класстардың атауы алынады.

```
SELECT DISTINCT ?term1 ?cls
WHERE {
?term1 a cs:term .
?res a swrc:Result .
?res swrc:isAbout ?term1 .
?res swrc:year 2020.
OPTIONAL { ?term1 a ?cls }
}
```

DISTINCT кілт сөзі қайтарылатын нәтижеден қайталанған көшірмелерді алып тастайды. Алдымен соңғы жылдан басқа барлық жылдардағы ғылыми еңбектің нәтижесіне сәйкес келетін мазмұнындағы барлық терминдердің (қайталануы мүмкін) $Terms$ жиынын құрастырамыз.

```
SELECT DISTINCT ?term2
WHERE {
?term2 a cs:term .
?res a swrc:Result .
?res swrc:isAbout ?term2 .
?res swrc:year ?year .
FILTER { ?year < 2020}
}
```

Бағдарлама жасайтын тілдің көмегімен $Terms_3 = Terms_1 \setminus Terms_2$ жиынының әртүрін қолданамыз. $Terms_1$ жиыны үшін бірінші сұраныстың нәтижесінен алынған класстардың атауы бойынша $Terms_3$ жиынының элементтерін топқа бөлеміз. Осы элементтердің тізімі қойылған сұраққа жауап ретінде беріледі.

Қызықтыратын бағыттың әдісі қолданылатын тапсырмалар тізбегі. Бұл сұранысты төмендегідей қарастырып көрейік. Біріншіден, берілген $T = \{t_1, \dots, t_n\}$ бағыттың барлық әдістерін ерекшелеп алу керек. Ол үшін әдіс болып

табылатын T терминдерінен method барлық түбірінің жиынын алу керек. Ары қарай осы әдістер орындалатын барлық тапсырманы таңдап алу керек, яғни $isSubTerm$ қатынасымен байланысты task экземпляр класын. Оны SPARQL тілінде формаландыру. Көп нүкте белгісін {UNION ?method isSubTerm t_i } түріндегі фрагментке ауыстыру керек, мұнда $i = 3, \dots, n-1$.

```
SELECT DISTINCT ?task
WHERE {
?task a cs:task .
?method a cs:method .
{ ?method isSubTerm t_1 }
UNION { ?method isSubTerm t_2 }
...
UNION { ?method isSubTerm t_n } .
?method isSubTerm ?task .
}
```

$isSubTerm$ байланысы онтологияда транзитивт ретінде анықталады. Бұл факт логикалық қорытындыны қолданғаннан кейін онтологияның t элементі $isSubTerm$ бірге турасынан байланысты болады. Логикалық қорытындыны қолдану сұраныстың жауабын қамтамасыз етеді.

Ғалымның ғылыми қызығушылықтарына байланысты тізім. Бұл сұранысты былай деп жазайық: ғылыми еңбектердің нәтижесімен байланысты барлық терминдердің тізімін көрсету. Ары қарай SPARQL тілінде бұл сұраныстың формальды жазбасы берілген.

```
SELECT DISTINCT ?term
WHERE {
?term a cs:term .
?res a swrc:Result .
?res swrc:isAbout ?term .
?res dc:creator P .
}
```

Уақыт бойынша зерттеудің жеке бағытына зерттеушілердің қызығушылық динамикасы. Бұл сұранысты былай деп жазайық: жылдар бойынша топтастырылған соңғы 5 жыл ішінде $T = \{t_1, \dots, t_n\}$ берілген бағыт бойынша ғылыми еңбектің нәтижелер санын беру. Оны SPARQL тілінде формаландырамыз.

```
SELECT DISTINCT ?res ?year
WHERE {
?res a swrc:Result .
?res swrc:year ?year .
{ ?res swrc:isAbout t_1 }
UNION { ?res swrc:isAbout t_2 }
...
UNION { ?res swrc:isAbout t_n } .
FILTER ( ?year > 2006 && ?year < 2020 )
}
```

Осы сұраныстың нәтижесінде соңғы 5 жыл ішіндегі T бағыты бойынша барлық ғылыми еңбектің нәтижелерінің тізімі беріледі. Осыдан кейін қолданылатын тілдің көмегімен бұл деректерді жылдарға бөлу қажет болады. Бұл тапсырма жұмыстың стандартты функциясының көмегімен іске асады.

Қызықтыратын бағытқа арналған конференциялар тізімі. Бұл сұранысты былай деп жазайық: $T = \{t_1, \dots, t_n\}$ бағытымен байланысты конференциялар тізімі беріледі. Оны SPARQL тілінде формаландырамыз.

```
SELECT DISTINCT ?conf
WHERE {
?conf a swrc:Conference .
{ ?conf swrc:isAbout t_1 }
UNION { ?conf swrc:isAbout t_2 }
...
UNION { ?conf swrc:isAbout t_n } .
}
```

Берілген конференцияға ұқсас конференциялар тізімі. Бұл сұранысты былай деп жазайық: Conf конференциясын сипаттайтын терминдермен байланысты конференция тізімі беріледі. Оны SPARQL тілінде формаландырамыз.

```
SELECT DISTINCT ?c
WHERE {
?c a swrc:Conference .
?c swrc:isAbout ?term .
Conf swrc:isAbout ?term .
}
```

Қызықтыратын бағытқа жұмыс істейтін зерттеушілердің тізімі. Бұл сұранысты былай деп жазайық: $T = \{t_1, \dots, t_n\}$ бағытымен байланысты конференциялар тізімі беріледі. Оны SPARQL тілінде формаландырамыз.

```
SELECT DISTINCT ?person
WHERE {
?person a swrc:Person .
?res a swrc:Result .
?res dc:creator ?person .
{ ?res swrc:isAbout t_1 }
UNION { ?res swrc:isAbout t_2 }
...
UNION { ?res swrc:isAbout t_n }.
}
```

Ұқсас тапсырмалармен айналысатын зерттеушілердің тізімі. Бұл сұранысты былай деп жазайық: терминдермен байланысты ғылыми зерттеулердің нәтижелер тізімі. Бұл сұраныста берілген P зерттеушінің ғылыми жұмысын сипаттайтын тізімді зерттеу терминдері байланысымен анықтаймыз. Оны SPARQL тілінде формаландырамыз.

```
SELECT DISTINCT ?person
WHERE {
?person a swrc:Person .
?res1 a swrc:Result .
?res1 dc:creator ?person .
?res1 swrc:isAbout ?term .
?term a cs:term .
?res2 a swrc:Result .
?res2 dc:creator P .
?res2 swrc:isAbout ?term .
}
```

Белгілі уақыт ішіндегі нақтылы бағыт бойынша жарияланымдар тізімі. Бұл сұранысты былай деп жазайық: Date=(startdate,enddate) мерзім ішіндегі $T = \{t_1, \dots, t_n\}$ берілген бағыттардың терминдерімен байланысты жарияланымдар тізімін береді. startdate мен enddate параметрлері кезеңнің бастауы мен аяқталу күндері болып табылады. Олар ЖЖЖЖ-АА-КК түріндегі жолдармен берілген. Оны SPARQL тілінде формаландырамыз.

```
SELECT DISTINCT ?pub
WHERE {
?pub a swrc:Publication .
{ ?pub swrc:isAbout t_1 }
UNION { ?pub swrc:isAbout t_2 }
...
UNION { ?pub swrc:isAbout t_n } .
?pub swrc:date ?date.
FILTER (?date > startdate^^xsd:date && ?date < enddate^^xsd:date)
}
```

Берілген жарияланымға ұқсас мақалалардың тізімі. Бұл сұранысты былай деп жазайық: Pub берілген жұмысын сипаттайтын терминдермен байланысты жарияланымдардың тізімін беру. Оны SPARQL тілінде формаландырамыз.

```
SELECT DISTINCT ?p
WHERE {
?p a swrc:Publication.
?term a cs:term.
?p swrc:isAbout ?term.
Pub swrc:isAbout ?term.
}
```

Қызықтыратын бағыттағы сұрақты шешудегі жеке ғалымның ғылыми үлесі. Бұл сұранысты былай деп жазайық: $T = \{t_1, \dots, t_n\}$ бағытының терминдерімен байланысты P ғалымның берілген ғылыми еңбегінің нәтижесімен байланысты жарияланымдардың тізімін беру. Оны SPARQL тілінде формаландырамыз.

```
SELECT DISTINCT ?res
WHERE {
?res a swrc:Result .
?res dc:creator P .
{ ?res swrc:isAbout t_1 }
UNION { ?res swrc:isAbout t_2 }
...
UNION { ?res swrc:isAbout t_n } .
}
```

Өткен кезеңге ғылыми ұйымның еңбек нәтижесіндегі жеке бөлімдерді ғылыми енгізу. Бұл сұранысты былай деп жазайық: Div бөлімінде жұмыс істейтін ғалымдардың Date=(startdate,enddate) уақыт мерзімінің ішіндегі ғылыми еңбектердің нәтижелерінің тізімі. Оны SPARQL тілінде формаландырамыз.

```
SELECT DISTINCT ?res
WHERE {
?res a swrc:Result .
?res swrc:date ?date .
?res swrc:creator ?person .
?person a swrc:Person .
Div swrc:member ?person .
FILTER (?date > startdate^^xsd:date && ?date < enddate^^xsd:date)
}
```

Осындай жолмен келесі бекітілген ережелерді қабылдауға мүмкіндік беретін SPARQL тілінде формаландырамыз.

1 Ереже. D ғылыми білім аймағы немесе оның пәндік аймақтағы мүмкін болған барлық терминдермен толықтырылған O_D онтологиясы мен олардың арасындағы байланыс берілді деп есептейік. Жеке ғалымдардың ғылыми

еңбектерінің нәтижелері жайлы деректермен толтырылған O_S ғылыми саладағы онтолгия берілді деп есептейік. Осы онтологиялардың арасында swrc:isAbout түрінің мүмкін болған барлық байланыстары орнатылған, яғни ғалымның ғылыми еңбегінің әрбір нәтижесіне оның тематикасын сипаттайтын көптеген белгілер бар. Сонда SPARQL сұраныс тілі мен O_D және O_S онтологияларының ұқсастығы сұраныстарына жауап алуға мүмкіндік береді.

Тізбектелген сұрақтарға жауап алу үшін белгілі талаптарға сай келетін ғылыми саланың O_S онтологиясы мен D білім аймағының O_D онтологиясының құрастырылуы керек.

SPARQL тіліндегі сұраныстар коды мен формальды үлгі арқылы жасалынатын жүйенің сұраныстарының арасындағы көрсетілген байланыс төменде келтірілген әсерлерді бақылауға мүмкіндік береді:

- бағдарламалық жүйе кодына қолданылатын онтологияның және сұраныстар жүйесінде (сонымен қатар жаңа түрінің де пайда болуы) қабылданған көптеген модификациялар;
- қарастырылатын сұраныстар мен қолданылатын онтологияға жүйе кодының бағдарламалық модификациясы.

Соңғы жағдай оның барлық өмірлік циклының сатыларындағы әсерлі верификациясы үшін қосымша мүмкіндіктер туғызады [41,с. 140].

2 бөлім бойынша тұжырым

Берілген бөлімде бес негізгі үлгіден тұратын ғылыми ақпаратты басқару жүйесі мен оның архитектурасына жасалған жалпы формальды үлгісі берілген. Деректерді жүктейтін үлгінің алгоритмдік негізі ретінде байланыс орнататын үлгілер пәндік аймақтағы онтология мен солардың арасында болады, сонымен қатар талап етілетін функционалдылықты алуға мүмкіндік беретін әдістер автордың көмегімен таңдап алынған, зерттеудің сәйкес келетін бағыты бойынша деректер көзінің қорытынды нәтижесі бойынша сұраныстарды орындайтын үлгілер бар. Бұл әдістерді іске асыратын бағдарлама құралдары жасалынған жүйеде қолданылу үшін сәйкестендірілген.

3 ТЕРМИНДЕРДІ ІРІКТЕП АЛУ АЛГОРИТМІ МЕН БІЛІМ АЙМАҒЫНДАҒЫ ОНТОЛОГИЯНЫ ҚАЛЫПТАСТЫРУ

Алдыңғы бөлімдегі негізгі тапсырманың бөлігі болып табылатын қосымша бес тапсырмалардың ішінен екеуін іріктеп аламыз. Оларды шешу үшін күрделі интеллектуалды алгоритмдерді құрастыру қажет. Соның ішіндегі біріншісі ғылыми білім деңгейі бойынша берілген аймақты сипаттайтын терминдерді ерекшелеп алуды мақсат етеді, ал екіншісі – осы терминдерді қолдану арқылы ғылыми білім аймағында онтологияны қалыптастыру болып табылады. Тапсырмаларды орындау үшін мәтінді лингвистикалық, статистикалық, семантикалық әдістерді, жекелей алғанда, Херст шаблондарының әдістемесін [42,с. 845] және іздеу жүйесінің көмегімен Интернеттен алынған іздеу нәтижелерінің мәтіндік қорытындысы қолданылады. Берілген бөлімде екі тапсырманың да шешілуіне негізделген алгоритмдердің сипаттамасы бар.

3.1 Берілген тематикалық бөлімдермен бірге мәтіндер жиынтығынан терминдерді іріктеп алу алгоритмі

3.1.1 Терминдерді іріктеп алу алгоритмінің сипаты

Doc құжаттар коллекциясында W – барлық сөздерің жиыны, ε – бос сөз, PW – реттелген сөздердің жұбы бар болсын, яғни $PW = W * W$. d құжаты берілген құжаттар жиынтығындағы n-ші бағытта тұрған сөздің әрбір натуралдық n санына сәйкес келетін $d: N \to W$ бейнелейді. Сөзі жоқ жағдайдағы нөмірлер (құжаттың соңынан кейін). Осыған сәйкес p жаңа жолы көрсетілген абзацта n-ші жағдайда тұрған n сөзінің әрбір натуралдық санына сәйкес келетін $p: N \to W$ көрінісі ретінде берілген. Сөз жазылмаған орынның номері бос сөз ретінде белгіленеді. Жиынтықтағы P арқылы барлық жаңа жолдарды белгілейміз. r айдарын шығаратын көптеген құжаттар ретінде, ал нақтырақ айтсақ – $r \in 2^{Doc}$ белгілейді. Айдардың қуаттылығын ондағы құжаттардың мөлшері сияқты $|r|$ арқылы белгілейтін боламыз. Берілген айдарлардың көпшілігін R арқылы белгілейміз.

Сонымен қатар, тағы да бірнеше қосымша белгілерді анықтаймыз:

- $\tau_1: PW \to W, \tau_2: PW \to W$ – жұп сөздер бірінші сөзге (соған сәйкес екінші) сәйкес келетін көптеген басқа сөздерге арналған жұптардың жобасы;
- $Freq: PW * Doc \to \mathbb{N} \cup \{0\}$ - $d \in$ Doc құжатында $p\varpi \in PW$ жұбының енгізілу санын анықтайтын функция;
- $Freq: W * Doc \to \mathbb{N} \cup \{0\}$ - $d \in$ құжатында $w \in W$ жұбының енгізілу санын анықтайтын функция;
- $L(d) = |\{n \in \mathbb{N} | d(n) \neq \varepsilon\}|$ - d құжатының ұзындығы;
- $id(a) = a$ – ұқсас бейнелеу;
- $Av(f, A) = \frac{\sum_{a \in A} f(a)}{|A|}$ - А соңғы жиын түріндегі f функциясының орташа мәні.

Мысалы, айдардағы құжаттардың орташа саны - $Av(|\cdot|, R)$, құжаттың орташа ұзындығы – $Av(L, Doc)$ – құжаттың орташа ұзындығы, $Av(id, A)$- $A =$

$\{a_1, \dots, a_k\}$ жиынтығының орташа арифметикалық саны.

Алгоритм үшін шығарылған деректер жиыны құжаттарда кездесетін, соған сәйкес сөздердегі кестеден тұрады. Алгоритмді қолданар алдында лемматизация - құжаттарға ең алдымен лингвистикалық өңдеу жұмысын жүргізген дұрыс болады, сөз формаларын дұрыс (сөздік бойынша) формаға келтіріп өңдеу керек. Мысалы, қазақ, орыс тіліндегі зат есім үшін бұндай форма жекеше түрдегі атау септігі болып табылады. Шығару кестесінің әрбір тармағында жазылған төрт сан бар: айдар нөмірі, құжат нөмірі, жаңа жол (абзац) нөмірі, сөз нөмірі. Егер А сөзі абзац бойында Б сөзінен бұрын кездесетін болса, онда кестеде де А сөзінің тармағы Б сөзінің тармағынан жоғары тұрады. Кесте алғашқы үш баған бойынша іріктелінген. Осындай жолмен, нақтылы сөз жайлы қандай құжатта, қанша рет, және қандай жерінде кездесетіні ғана белгілі.

Алгоритм төрт сатыдан тұрады. Олардың әрқайсысында кейбір ережелердің көмегімен алдыңғы қадам арқылы алынған M_i мен M_{i-1}, жиынтығы таңдалып алынады. Бірінші сатыда таңдау PW (сөздердің барлық жұптары) жиынтығынан алынып, іске асады, яғни $M_0 = PW$. M_4 жиынтығы термин – жұптар болып табылады, ол барлық төрт өлшемді де қанағаттандырады.

3.1.2 Кеңістікті өлшем

Келесі өңдеуге арналған жұптардың алғаш таңдап алынуы терминді білдіретін сөздер мәтіннің ішінде айтарлықтай жақын орналасқандығы (дегенмен, жақын орналасуы міндетті емес) жайлы болжамға негізделеді:

$$M_1 = \{pw \in M_0 | \exists p \in P: |p^{-1}(\tau_1(pw))| \leq MAX_{DIST}\}. \quad (3.1)$$

Мәтінде MAX_DIST-1 көп кездесетін басқа сөздердің арасындағы бір абзацта орналасқан екі сөз жұпты құрайды.

3.1.3 Жиілікті өлшем

Жоғарыда атап өткендей, алгоритм жұмысының нәтижесінде ерекшеленген сөздердің жұбын мәтіннің векторлы кеңістік базисі ретінде қарастыруға болады. Бұндай базис ақпараттық болу үшін онда мәтіннің ішінде көп кездесетін сөз болуы керек. Бұл талапты қамтамасыз ету мақсатында барлық жиынтықтарда MIN_FREQ қарағанда кем кездесетін M_1 жұптарының көбінен ерекшеленетін нақты өлшемдер қолданылады:

$$M_2 = \{pw \in M_1 | \sum_{r \in R} \sum_{d \in r} Freq(pw, d) \geq MIN\_FREQ\} \quad (3.2)$$

3.1.4 Сипаттамалық өлшем

Сипаттамалық өлшем – алгоритмнің негізгі өлшемі. Оның мәні терминді анықтауда жатыр, жұп кейбір айдар үшін сай келуі керек.

Айдардағы жұптың салмағы. Әрбір жұпқа сандар жиынтығы сәйкес келеді – айдардың әрқайсысындағы жұптардың салмағы. r айдарындағы pw жұбының салмағы келесі формула арқылы алынады:

$$Weight_\tau(pw) = \frac{\sqrt{\sum_{d \in r} \ln\left(\frac{\sqrt{Freq(pw,d)}}{\ln\left(\frac{r}{Av(L,Doc)}+1\right)}+1\right)}}{\ln\left(\frac{r}{Av|\cdot|,R}\right)} + 1 \qquad (3.3)$$

Осыған сәйкес $Weight_\tau(pw)$ айдарындағы сөздің салмағы анықталады, нақтырақ айтсақ – жоғарыда берілген формулада $Freq(pw,d)$ белгісі $Freq(w,d)$ ауысады.

Салмақ функциясын таңдау. r айдарында жұптардың салмағын $Weight_\tau(pw)$ функция түрін таңдаудың негізін анықтау арқылы көрсетеміз. r айдары k құжаттарынан $d_1, \dots, d_k$ тұрады делік. $k > 1$ деп есептейік.

Қойылатын талаптар.

Ең алдымен айдардағы сөздердің жұптарының қызметіне сәйкес келетін талаптарды таңдап аламыз (осыған сәйкес басқа да жағдайлар кездеседі):

- айдардағы жұптардың салмағы – қарсылықты емес, қолданыстағы сан;
- құжатқа енетін жұптардың салмағы артқан кезде айдардағы жұптың салмағы артады;
- құжатқа қатысты ұзындық артқан кезде айдардағы жұптың салмағы азаяды;
- Құжаттарға қатысты сан артқан кезде айдардағы жұптардың салмағы азаяды;
- Егер жұп k құжатында n рет кездесетін болса, онда оның айдардағы салмағы құжатта кездескен жағдаймен салыстырғанда (айдардағы құжаттардың барлығы бірдей ұзындықта болса) артады. Қойылатын талапта айдардағы құжаттарға қатысты сандар мен құжаттың ұзындығы ескеріледі. Бұл айдардың орташа қуаты мен құжаттың орташа ұзындығының көрсеткіші әртүрлі болатын құжаттардың әртүрлі жиынтығынан жұптардың салмағын салыстыру мүмкін болу үшін жасалынған. Тізімдегі талаптардың соңғысы сөздердің жұптары айдардың көптеген құжаттарында кездесетін болса, оның бұл айдар үшін маңыздылығы айдар құжатына енетін жұптардың енетін мөлшері біреу ғана болған жағдаймен салыстырғанда сол мәннің қысқаша қалыптасқан жиынтығы болып табылады.

Параметрлер. салмақ қызметінің параметрлері:

- айдардың құжатына енетін жұптардың саны: $Freq(pw, d_i), i \in \overline{1,k}$ F;
- айдардың құжаттарының ұзындық мөлшері: $\frac{L(d_i)}{Av(L,Doc)}, i \in \overline{1,k}$;
- айдардағы құжаттар санының мөлшері: $\frac{r}{Av|\cdot|,R}$.

Соның нәтижесінде алатынымыз $Weight_r$ функциясы $2k+1$ параметрлеріне тәуелді болуы керек:

$$Weight_\tau(pw) =$$
$$= Weight_\tau\left(Freq(pw,d_1),\dots,Freq(pw,d_k),\frac{L(d_1)}{Av(L,Doc)},\dots,\frac{L(d_k)}{Av(L,Doc)},\right) =$$
$$= Weight_\tau(x_1,\dots,x_k,y_1,\dots,y_{k,z}) = (Weight_\tau(\tilde{x},\tilde{y},z).$$

$Weight$ функциясындағы r индекс $Weight_\tau$ құжаттарының әртүрлі санынан тұратын айдар үшін арналған фактіні меңзейді, функциялар аргументтердің әртүрлі санына байланысты болады. $Weight_\tau$ функциясын анықтаудың аймғын ескереміз:

$$D(Weight_\tau): x_i \in \mathbb{Z} \cap [0,+\infty), y_i \in (0,+\infty), z \in (0,+\infty), i \in \overline{1,k} \qquad (3.4)$$

Формалды талаптар. Енгізілген белгілерді қолдану арқылы $Weight_r$ функциясына қойылатын талаптарды қалыптастырамыз. Олар төменде тізбектелген.

1. Қарсылықты емес мән: $Weight_\tau(pw) \geq 0$.
2. Жиілікке тура тәуелділік: $x_i^{'} > x_i$ болған кездегі $\forall i \in \overline{1,k}$
$$Weight_\tau(x_1,\dots,x_{i-1},x_{i+1},\dots,x_k,\tilde{y},z).$$
3. Құжаттың ұзындығына тікелей тәуелділік: $y_i' > y_i$ болған кездегі $\forall i \in \overline{1,k}$
$$Weight_\tau(\tilde{x},y_1,\dots,y_{i-1},y_i',y_{i+1},\dots,y_k,z) <$$
$$< Weight_\tau(\tilde{x},y_1,\dots,y_{i-1},y_i,y_{i+1},\dots,y_k,z).$$
4. Айдардың қуаттылығына байланысты кері тәуелділік: $z' > z$ болғанда
$$Weight_\tau(\bar{x},\bar{y},z'), < Weight_\tau(\bar{x},\bar{y},z).$$
5. Жұптар кездесетін құжаттардың санына тура тәуелділік:
$y_1 = y_2 = \dots = y_k, \forall n \in \forall l \in \overline{1,k}$ болғанда

$$Weight_\tau\left(\underbrace{n,\dots,n}_{k},\tilde{y},z\right) > Weight_\tau(\underbrace{\overbrace{0,\dots,0,kn}^{l},0,\dots,0}_{k},\tilde{y},z) \qquad (3.5)$$

Қызметтің жалпы көрінісі. Қалыптасқан талаптарды қанағаттандыратын қызметтерді іріктеп алу үшін $Weight_r$ функциясының келесі жалпы көрінісі таңдап алынады:

$$Weight_r(\bar{x},\bar{y},z) = h\big(g\big(f(x_1,y_1),\dots,f(x_k,y_k)\big),z\big) \qquad (3.6)$$

$f(x,y)$ функциясы құжаттағы жұптардың салмағын анықтайды. Салмақты анықтауда жұпта $g(x_1,\dots,x_k)$ функциясы қолданылады. $h(x,y)$ функциясы арқылы кейін айдарға қатысты қуаттан айдардағы жұптардың салмағының тәуелділігін көрсетіледі.

$$Weight_r(\bar{x}, \bar{y}, z) = f\big(g\big(f(x_1, y_1), \dots, f(x_k, y_k)\big), z\big) \qquad (3.7)$$

f мен g функциялары төмендегідей анықталу жолы мен мәніне ие болуы керек:

$$D(f): x \in [0, +\infty), y \in (0, +\infty), E(f) = [0, +\infty),$$
$$D(g): x_i \in [0, +\infty), i = \overline{1, k}, E(g) = [0, +\infty) \qquad (3.8)$$

1-4 талаптардың орындалу үшін $Weight_\tau$ функциясына қатысты теорема қалыптастырылды.

1 лемма. $f = f(x, y)$ мен $g = g(x_1, \dots, x_k), k > 1$ g функциясы жоғарыда көрсетілгендей мән мен анықталу жолынан тұрады делік. Осы функциялардың барлық аймағындағы келесі шарттар орындалды делік:

1. $x' > x$ кезінде $f(x', y) > f(x, y)$;
2. $x' > x$ кезінде $\forall i \in \overline{1, k}\ g(x_1, \dots, x_{i-1}, x_i', x_{i+1}, \dots, x_k) > g(x_1, \dots, x_{i-1}, x_i, x_{i+1}, \dots, x_k)$
3. $y' > y \qquad f(x', y) > f(x, y)$;

Сонда 1-4 функцияларына қойылатын талаптар $Weight_\tau(\bar{x}, \bar{y}, z) = f(g\big(f(x_1, y_1), \dots, f(x_k, y_k)\big), z)$ орындалады.

Дәлел. Салмақтың қарсылықты болмауы f функциясының мәні арқылы пайда болады. Жиілікке тура тәуелділік $x_i' > x_i$болған кездегі 1 мен 2 шартынан шығатын жиілік пайда болды, соған сәйкес $f(x', y) > f(x, y)$ кезінде $g\big(f(x_1, y_1), \dots, f(x_i', y_i), \dots, f(x_k, y_i)\big) > g\big(f(x_1, y_1), \dots, f(x_i, y_i), \dots, f(x_k, y_k)\big)$, олай болса, $f\big(g\big(f(x_1, y_1), \dots, f(x_i', y_i), \dots, f(x_k, y_i)\big), z\big) > > f\big(g\big(f(x_1, y_1), \dots, f(x_i, y_i), \dots, f(x_k, y_k)\big), z\big)$. Құжаттың ұзындығынан кері тәуелділік 1, 2 мен 3 шартынан шығады: $y_i' > y_i\ \ f(x_i, y_i') > f(x_i, y_i)$ болған кезде, осыған сәйкес, $g\big(f(x_1, y_1), \dots, f(x_i, y_i'), \dots, f(x_k, y_k)\big) < g\big(f(x_1, y_1), \dots, f(x_i, y_i), \dots, f(x_k, y_k)\big)$, олай болса, $f\big(g\big(f(x_1, y_1), \dots, f(x_i, y_i'), \dots, f(x_k, y_k)\big), z\big) < < f\big(g\big(f(x_1, y_1), \dots, f(x_i, y_i), \dots, f(x_k, y_k)\big), z\big)$. Айдардың қуаттылығына кері тәуелділік 3 шарт арқылы іске асады.

Тестілеу. Тестілеу үшін төмендегідей функциялар таңдалып алынған:

- $f(x, y)$:
  1. $\frac{\sqrt{x}}{\ln(1+y)}$;
  2. $\frac{x^2}{y}$;
  3. $\frac{x\sqrt{x}}{y}$;
  4. $\frac{x(1+\ln(1+x))}{y}$;
  5. $\frac{xln(1+x)}{y}$;

6. $\frac{x}{\ln(1+y)}$;
7. $\frac{x}{y}$;

- $g(x_1,\ldots,x_k)$;

. $\sum_{i=1}^{k} \ln(1+x_i)$;

. $\prod_{i=1}^{k} (1+x_i)$;

. $(\sum_{i=1}^{k} x_i) \cdot |\{x_i | x_i > 0\}|$.

Бұл функциялардың барлығы 1 теореманы қанағаттандырады, ол дегеніміз салмақ функциясына деген 1-4 талаптардың барлығын қамтамасыз ететіндікті білдіреді. Тестілеу нәтижесі бойынша алынған функциялар комбинациясы осы талаптардың орындалуын қамтамасыз етеді.

Тестілеуді жүргізуді бақылау үшін аргументтердің аздаған санына тәуелді болып келетін аса қарапайым түрінің салмақтық 3 функциясы қолданылады.

$$Weight_\tau(pw) = \sum_{i=1}^{k} Freq(pw, d_i);$$

$$Weight_\tau(pw) = \sum_{i=1}^{k} \ln(1 + Freq(pw, d_i));$$

$$Weight_\tau(pw) = (\sum_{i=1}^{k} Freq(pw, d_i)) \cdot |\{d \in r | Freq(pw, d) > 0\}|.$$

Деңгей бойынша тестілеу f мен g функцияларының кейбір комбинациялары бойынша өткізілген. Оның нәтижесі бойынша, біліктіліктің ең жоғарғы деңгейі функцияның келесі комбинациясын қолданған кезде ғана қол жетімді болады:

$$f(x,y) = \frac{\sqrt{x}}{\ln(1+y)}; g(x_i, \ldots, x_k) = \sum_{i=1}^{k} \ln(1+x_i) \qquad (3.9)$$

Жоғарыдағы ақпаратты ескеру арқылы, айдардағы жұптың салмағының қызметі ретінде келесі функция таңдап алынған:

$$Weight_\tau(pw) = \frac{\sqrt{\sum_{d \in r} \ln\left(\frac{\sqrt{Freq(pw,d)}}{\ln\left(\frac{L(d)}{Av(L,Doc)}+1\right)}+1\right)}}{\ln\left(\frac{|r|}{Av|\cdot|,R}+1\right)} \quad (3.10)$$

Айдардағы жұптардың салмағын анықтау үшін белгілі $Weight_\tau$ функциясын қолдануға мүмкіндік беретін теореманы дәлелдейік.

1 теорема. Таңдап алынған $Weight_\tau(pw)$ функциясы талаптарға сай келеді. Дәлелдеме. 1 леммадағы f мен g функцияларының қасиетінен 1-4 талаптарының орындалуы туындайды.

$k \in \mathbb{N}, k > 1, y_1 = \ldots = y_k = y_0, n \in \mathbb{N}, l \in \overline{1,k}.$ болсын. Сонда

$$Weight_\tau\left(\underbrace{n, \ldots, n}_{k}, \bar{y}, z\right) = f\big(g\big(f(n,y_0), \ldots, f(n,y_0)\big), z\big) = f\big(k \ln\big(1+f(n,y_0)\big), z\big).$$

Сонымен қатар

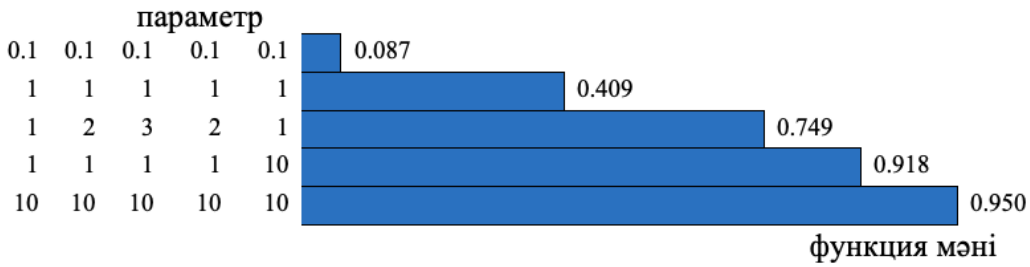
$$Weight_\tau\left(\underbrace{\overbrace{0, \ldots, 0}^{l}, kn, 0, \ldots, 0}_{k}, \bar{y}, z\right) = f(g(\overbrace{0, \ldots, 0}^{l}, f(kn,y_0), 0, \ldots, 0), z) = f(\ln(1+f(kn,y_0)), z).$$

f функциясының өсу күшіндегі бірінші аргументке келесі формулаларды келтіруге болады, $kln\big(1+f(n,y_0)\big) > \ln(1+f(kn,y_0))$ немесе, дәл соны білдіретін $\big(1+f(n,y_0)\big)^k > 1+f(kn,y_0)$. Сонымен қатар $f(x,y) = \frac{\sqrt{x}}{\ln(1+y)}$ функциясы $k \in \mathbb{N}, k > 1, n \in \mathbb{N}$ кезіндегі $f(kn,y) < kf(n,y)$ теңсіздігін қанағаттандырады. Демек, $1+f(kn,y_0) < 1+kf(n,y_0)$. Алайда түсінікті жағдай $(1+f(n,y_0))^k > 1+kf(n,y_0)$, себебі $f(n,y_0) > 0$. Дәлелдеу осыны қажет еткен.

Жұптардың сипаттамалық функциясы. Айдардағы жұптардың есептеліп алынған салмағына бірдей уақытта нольге тең болмайтын $A = \{a_1, \ldots, a_n\}$, қарсылықты емес $Discr$, сипаттамалық функциясы қолданылады, сан – сипаттамалық көрсеткіші болып табылады:

$$Discr(a_1, a_2, \ldots, a_n) = 1 - \frac{Av(id,A)}{max_{a \in A} \cdot (1 + \ln(1 + max_{a \in A} a))} \quad (3.11)$$

Бұл функция [0, 1] кесінділерінің мәндерін қабылдайды, және терминді анықтау жағынан алып қарағанда бұл қаншалықты сипаттамаға ие болса, соншалықты бұл функцияның мәні 1 функциясына жақын келеді. 3.1 суретте параметрлердің үлгілері мен ондағы $Discr$ функциясының мәні көрсетілген.



Сурет 3.1 - $Discr$ сипаттық функциясының нәтижелері

$$R = \{r_1, \dots, r_n\} \tag{3.12}$$

Егер

$$M_3 = \{pw \in M_2 | Discr(Weight_{\tau 1}(pw), \dots, Weight_{\tau n}(pw) \geq MIN\_DISCR\}.$$

Осындай жолмен [0, 1] кесіндісіне жататын MIN_DISCR константт аз болмайтын $Discr$ функциясының мәні бар жұптар таңдап алынады.

3.1.5 Мәні бар айдарлардың өлшемі

Мәні бар айдарларда термин деп саналатын сөздер жұптары жиі кездесу керек. Егер осы айдарға жататын сөз кездессе, онда сол сөзбен сипаттық өлшемді қанағаттандыратын басқа сөздердің кашықтығы аз болуы керек. Яғни ол бұл сөз жұптары басқа айдарда кездеспейді, жиын термині болып табылмайды.

3.1 анықтама. pw жұптары үшін r айдары мәнді болады, егер

$$Weight_\tau(pw) \geq Av(Weight_\cdot(pw), R) \tag{3.13}$$

$2^R = \{A | A \subseteq R\}$ болсын. Ол үшін айдардағы мәні бар әрбір жұпқа сәйкес келетін $M_3 \to 2^R$ көрсеткішін анықтаймыз:

$$imp(pw) = \{r \in R | Weight_\tau(pw) \geq Av(Weight_\cdot(pw), R)\} \tag{3.14}$$

Сонда

$$M_4 = \{pw \in M_3 | \min_{r \in imp(pw)} \left(\frac{Weight_\tau(\tau_1(pw))}{Weight_\tau(pw)} + \frac{Weight_\tau(\tau_2(pw))}{Weight_\tau(pw)}\right) \leq MAX\_FREQ\_RATIO\} \tag{3.15}$$

Осындай жолмен, олар үшін мәні бар айдарлардың ішінен айдар табылатын жұптар іріктеп алынады, оның ішіндегі әр сөздердің жұбының салмағы жұптың салмағынан MAX_FREQ_RATIO есе артық болады.

Терминдерді ерекшелеу алгоритмінің есептелінетін қиындығын бағалауға мүмкіндік беретін теореманы құрастырамыз және дәлелдейміз.

2 теорема. R жиынынан айдарға бөлінген Doc құжаттарының жиыны берілді делік. W – осы құжаттың ішінде кездесетін барлық әртүрлі сөздердің жиыны делік, ал $m = |W|$ – осындай сөздердің саны. $L(d)$, $d \in Doc$ – құжаттың

ұзындығын білдіретін функция деп алынсын (құжаттың ішіндегі сөздердің саны). $n = \sum d \in Doc\ L(d)$ – құжаттың жиынындағы барлық сөздердің саны деп белгілейік. Сонда алгоритм үшін терминдерді ерекшелеп алу үшін келесі бағалау түрі әділ болмақ:

- алгоритмнің уақытша күрделілігі (нашар жағдайда):

$$O(|R||Doc|\min(m^2, n) + n)$$

- кеңістік бойынша күрделілігі (нашар жағдайда):

$$O(|Doc|\min(m^2, n))$$

- жұмыс нәтижесінен алынған терминдердің саны келесіге тең:

$$O(\min(m^2, n))$$ тең.

Алгоритмнің уақытша күрделілігі тапсырма қатарларын шешуде орындалатын қарапайым операциялары (арифметикалық және салыстыру операциясы) үшін керек. Кеңістік қиындығының алгоритмнің жұмысы үшін ерекшелеп алуға қажетті жадының ұяшықтары үшін қажет.

Дәлелдеме. Терминдерді ерекшелеп алу алгоритмі төрт қадамнан тұрады. Олардың әрқайсысына қорытынды жасайық. Кеңістік өлшемі бойынша жұптар жиынын таңдау үшін барлық құжаттар жиынының өтуге арналған бір жолы қажет. Әрбір сөз үшін рет-ретімен қолданылатын сөздердің келесі MAX_DIST қолданылады, және осы жұптар M_1 жиынына қосылады. Сонымен қатар әрбір жұп үшін әрбір документке қатысты енгізілу саны есте сақталынады. Осындай жолмен, бұл қадамның уақытша күрделілігі $O(n)$ тең болмақ. Ыңғайлы болуы үшін $k = |M_1|$. белгілейміз. $k \leq MAX\_DIST \cdot n$ және $k \leq m^2$ жиын жұптары үшін әрбір жұпқа MAX_DIST аспайтын жұптар қосылады, ал екіншісі – себебі, $m_2 - m$ қуат жиынының ішіндегі екі сөзден алынған реттік комбинациялардың барлығының саны. Бірінші қадамның кеңістік бойынша күрделігі $k|Doc|$ тең, себебі әрбір жұп үшін әрбір құжатқа енгізілу санын сақтау аса қажет болып табылады.

Екінші қадамда M_1 жиынынан MIN_FREQ реттен кем болмайтын колекцияның барлық құжаттарында кездесетін жұптар алынып тасталады. Сондықтан жұптарды іріктеп алу уақыты $O(k)$ тең. Екінші қадамның кеңістік бойынша күрделілігі $O(1)$ тең, себебі қосымша ешқандай ақпарат сақталынбайды. Жұп не есте қалады, болмаса алып тасталынады. Ескере кетейік, екіншіден төртіншіге дейінгі алгоритмнің сатыларында бірінші сатыда іріктеп алынған жұптарды іріктеу жүзеге асырылады. Сондықтан $|M_4| \leq |M_3| \leq |M_2| \leq |M_1| \leq k$.

Жұптардың әрқайсысы үшін үшінші қадам кезінде алдыңғы қадамда іріктелініп алынған жұптардың M_2 жиыны айдардың әрқайсысының салмағына қарай есептелінеді. pw жұп пен r айдары үшін $O(|r|)$ тең операцияның есебі орындалады, себебі есептеу барысында айдардың әрбір құжатына жұптардың енгізілу жиілігі орын алады. Осындай жолмен, салмақты есептеу кезіндегі операцияның жалпы саны $O(k|R||Doc|)$ тең. Жұптардың сипаттық қызметін анықтау үшін $O(|R|)$ аспайтын операцияларды орындау қажеттілігі туындайды. Осылай, бұл сатының жалпылама уақытша күрделілігі $O(k|R||Doc|)$ тең. Ұзақ мерзімді k тестіленетін жұптардың жалпы санына баға береді.

Үшінші қадамның кеңістіктік күрделілігі $O(|R|)$ тең . Бұл баға келесі болжам арқылы туындайды. Бірінші қадамға тексерістен өткен жұптар ғана енетін алгоритмнің 2-4 қадамдары ешбір жұпқа тәуелді болмай орындалуы мүмкін. Оның мәні сонда, M_1 жиынтығы мен келесі анықтау жолы арқылы ары қарай жұптарды таңдауда жұптардың санына тәуелді болмайтын жадының көлемі M_4 жиынының қорытындысы мен M_2, M_3 жиынына түседі. Осы жерден $O(|R|)$ пайда болады.

Әрбір жұп үшін төртінші қадамда мәні бар көптеген айдар анықталады. Осы кезде барлық жұптар үшін барлық айдарда салмақтың мөлшері анықталады, сондықтан жұптардың мәні бар айдарының жиыны үшін операциялардың саны $O(|R|)$ тең. Әрбір жұп үшін өлшемнің сәйкес келуін тексеру үшін жұпты құрайтын екі сөздің салмағын анықтау керек, бұл әрбір мәні бар айдарға қатысты. Соған сәйкес үшінші қадам үшін соған қажетті орындалатын $O(|R||Doc|)$ операциясын орнымдаймыз. $O(k|R||Doc|)$ операциясы және k тестіленетін жұптардың жалпы мөлшеріне бағаланады.

Алгоритмнің төртінші қадамының кеңістіктік күрделілігі $O(|R|)$ тең. Талап етілетін жадының жалпы көлемі қолданылатын айдарлардың санына байланысты.

Алгоритмнің барлық төрт сатысы үшін бағалауды біріктіре отырып оның жалпы уақытқа байланысты күрделілігі $O(n) + O(k) + O(k|R||Doc|) + O(k|R||Doc|) = O(n + k|R||Doc)$ тең деп аламыз. k қолданып келесі формуланы аламыз $O(|R||Doc|\min(m^2, n) + n$. Осыған сәйкес алгоритмнің кеңістіктік күрделілігі келесідей болады $O(k|Doc|) + O(1) + O(|R|) = O(k|Doc|) = O(|Doc|\min(m^2, n))$.

Терминдер мөлшері: $t = |M_4| \leq |M_1| = k = O(\min(m^2, n))$ (оны t арқылы белгілейміз). Осы жерден $t = O(\min(m^2, n))$.

Ескерту. $\min(m^2, n)$ белгісінде m^2 мен n мәндері әртүрлі қатынастарда болуы мүмкін. Мысалы, әртүрлі құжаттарда 10 сөзден кездесетін 100 түрлі $(100^{10} = 10^{20})$ сөз бар. Онда $\mathrm{m} = 100$, $\mathrm{m}^2 = 10^4$, ал $\mathrm{n} = 10^{21}$, мен $\mathrm{m}^2 \ll n$. Яғни m_2 сөзден m_1 құжаттарында кездесетін $\mathrm{m} = m_1 m_2$ әртүрлі сөздер берілген болсын. Сонда m үлкен болғандағы $n = m_1 m_2 = m \ll m^2$.

3.2 Ғылыми білім аймағындағы онтологияның құру алгоритмі

Білім аймағындағы онтологияның құрылу алгоритмі төменде көрсетілген жеті сатыдан тұрады.

- N_C^D түсінігінің атауларының жиынынан түзілу.

Бұл жиын ғылыми білімнің әртүрлі аймақтары үшін жалпы концептілерден тұрады.

- D ғылыми білімнің берілген аймағын сипаттайтын терминдерді ерекшелеу

Бұл сатыда терминдерді ерекшелеп алу алгоритмі конференцияның анонстық коллекциясынан кілт сөзін бөліп алады.

- Терминдерді фильтрлеу.

Фильтрлеу сатысында алынған терминдердің сапасы терминнің қаншалықты тұрақты белгі екендігі бағаланады, ал екіншіден, термин D ғылыми білім аймағы

мен терминнің арасында қаншалықты байланыс бар екендігі анықталады. Бекітілген өлшемге сай келмейтін терминдер алып тасталынады.

- Терминдер арасындағы ассоциациялық байланысты ерекшелеу.

Бұл сатыда салыстырмалы түрде бірдей веб-парақшаларында кездесетін терминдердің жұптары таңдап алынады. Бұл қасиетті тексеру үшін Интернет желісіндегі іздеу жүйесінің сұраныс қорытындысы қолданылды.

- Терминдер иерархиясын орнату.

Алдыңғы сатыда араларына байланыс орнатылған термин жұптарының әрқайсысы үшін Херст лингвистикалық шаблон модификациялық әдісі қай иерархиялыққа жатындығы тексерілді.

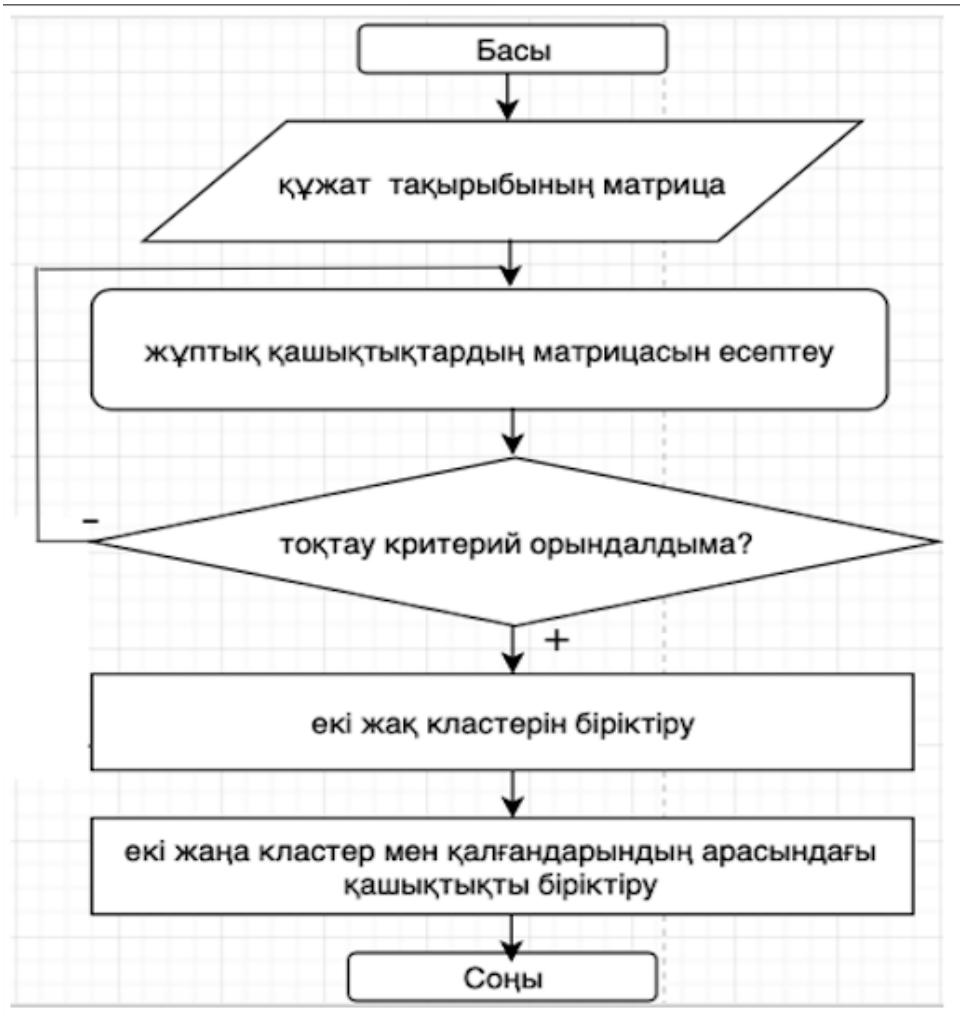
- Терминдерді қазақ, орыс тіліне аудару.

Онтологияға терминдерді қазақ, орыс тілінде енгізу үшін Википедиядағы ағылшын тіліндегі мақалалардың орысша, қазақша баламаларын қолданып автоматты түрде қосады.

- терминдерді онтологиялық түсінік бойынша саралау.

Осы қадамда алынған термндер талданылады. Байланысын анықтау үшін Интернет ізденіс жүйесіне сұраныстар нәтижесі қолданылады.

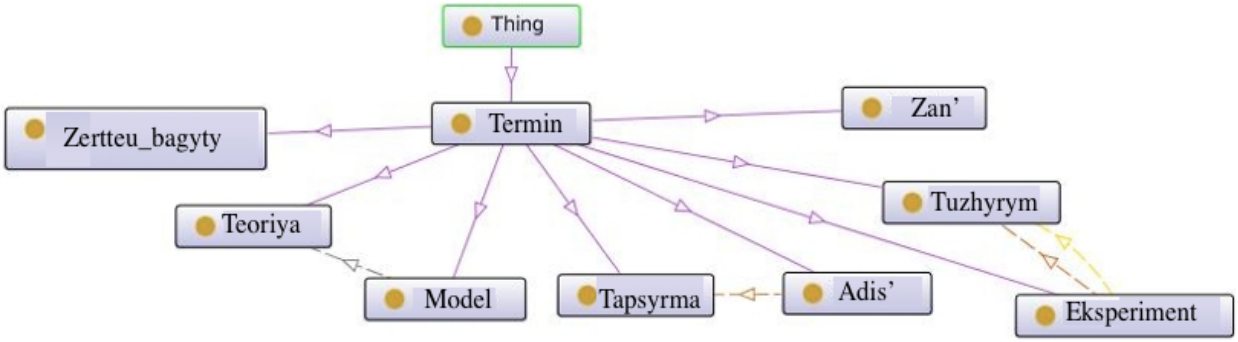
Онтологияны құрастырудың алгоритмі сурет 3.2 берілген.



Сурет 3.2 – Ғылыми білімнің жеке аймағындағы онтологияны құрастырудың алгоритмдік сызбасы

3.2.1 Түсініктердің көптеген атауларының құрастырылуы

60 элементтен тұратын N_C^D онтологиялық түсінік атауларының тізімі ғылыми білімнің барлық аймағына бірдей. Құрамында парадигма, әдіс, түсінік, бағыт, алгоритм және олардың ағылшын тіліндегі эквиваленттері кіреді. N_C^D түсінігінің атауларының фрагменті сурет 3.3 бейнеленген.



Сурет 3.3 – Ғылыми білім аймағындағы онтология түсінігі (фрагмент)

3.2.2 Терминдерді ерекшелеу алгоритмі

Конференция анонс жиынынан алгоритим көмегімен терминдерді ерекшелеп алу сатысында кілт сөздер анықталады. Алдымен құжаттар айдарға бөлінеді. Конференция анонсында білімнің әртүрлі аймақтарына бағытталған сілтемелер тізімін қолданады. Құрастырылып жатқан онтологияның D білім аймағына айдардың біреуі сәйкес келуі керек. Бұл айдарды тұтас деп атаймыз. Бұндай жағдайда конференцияның анонсы каталогта айдарға бөлінбеген фактімен байланысты аздаған қиындықтар туындайды. Оның орнына оларды белгі арқылы ерекшелейді. Терминдерді ерекшелеу алгоритмін бұл құжаттарға қолдану үшін белгінің көмегімен құжаттарды айдарға бөлу қажет.

$Docs$ (ғылыми конференция анонсы) мәтіндік құжаттар жиынын (call for papers), осы конференциялардың тақырыптарын сипаттайтын $Tags$ мәтіндік белгінің жиынын қарастырамыз. $EventTags: Docs \to 2^{tags}$– көптеген белгілері бар құжаттарға сәйкес келетін белгі деп алайық. $R^{||Tags||}$ векторлық кеңістігін қарастырамыз, оның базистік векторлары $Tags$ жиынының элементтеріне сәйкес. Осы векторлық кеңістік нүктесіндегі $Docs$ жиынынан құжатты алып белгілейміз. Ол үшін әрбір құжатқа құжаттағы i номерімен белгінің енуіне сәйкес келетін i -ші координатының $||Tags||$, ұзындық бірлігі мен нөлден теремiз. Барлық вектор-құжаттарындың координасы нөлге, не бірге тең, сондықтан бір құжатта бір тег белгісі болады.

Жиі қолданылатын орташа k кластеризация алгоритмін тәжірибеден өткізе келе, кластер-айдарларға векторлы кеңістік нүктесін бөлеміз. Нүктелер арасында $d(u,v)$ қашықтығының метрикасы есебінде m ұзындығының u, vәртүрлі векторларының стандартты функциясы қолданылады:

$$d(u,v) = \frac{c_{01}+c_{10}}{m} \quad (3.16)$$

мұнда $c_{ij} - u_k = i, \quad v_k = j$ болатын $(u_k, v_k), k \in \{\overline{1,m}\}$ біркелкі индекспен элементтердің жұп саны. Кластеризациялау нәтижесінде әрбір құжат тура бір айдарға келіп түседі. Айдар жиынын $R = \{r_1, \dots, r_n\}$ арқылы белгілейміз.

Сондағы көптеген белгілерден тұратын әрбір айдарға белгі қоямыз. $r \in R$ айдарын аламыз. r айдарынан сәйкес құжаттардың барлық белгілерінен тұратын $Tars_r$ жиынын құраймыз. $Tars_\tau$ жиынынан әрбір tag белгі үшін TF-IDF классикалық формуласын қолдану арқылы берілген айдарда оның салмағын анықтаймыз W [49]:

$$W(tar_\tau R) = \frac{n_{tag}}{N_\tau} * \frac{n}{\left|\{r_j \in R: tag \in Tags_{rj}\}\right|} \tag{3.17}$$

Мұндағы $n_{tag} - tag$ белгісі бойынша r айдарынан ерекшеленген құжаттар саны; N_r – қайталану жағдайы ескерілген r айдарынан сәйкес келетін құжаттардың белгілерінің саны. Бұл формуланың мәні сонда, айдардағы белгілердің мөлшері осы белгі арқылы ерекшеленген сондағы құжаттардың санына тура пропорционал, және айдардағы белгілердің жалпы санына кері пропорционал. r айдарының бейнесі ретінде осы айдардағы салмағы жоғары C_p белгісінен таңдап аламыз. C_p саны алгоритмнің параметрі болып табылады.

Барлық айдардың портретін құрастырғаннан кейін D білімінің аймағындағы осы портреттерді қолмен қарап шығуы керек, және D білім аймағы портретіне тура сәйкес келетін $r_D \in R$ толықтай айдарын көрсету керек. Егер де эксперттің пікірі бойынша айдардың ешқайсысы D сәйкес келмейтін болса кластерлердің талап етілетін мөлшерін өзгерту арқылы кластерилизацияны қайта жүргізуі керек. Егер де айдарлардың көбінің портреттері D аймағына жауап беретін болса бұл айдарлар біреуіне ғана тоғысу үшін кластерлердің санын азайту керек болады.

Айдарларға бөлінген құжатқа терминдерді ерекшелеп алу алгоритмі қолдалады. Алгоритм жұмысының нәтижесі $Terms$ терминдер жиынында болады. Біз одан r_D толық айдарындағы құжатта бір рет болса да кездесетін терминдерден тұратын $Terms_1 \subseteq Terms_2$ жиын ішіндегі жиынды ерекшелейміз. Айдарлардағы терминдердің салмағы терминдерді ерекшелеп алу алгоритмдерінің жұмысы барысында анықталады. Алынған $Terms_1$ жиыны D білімінің айдынындағы конференцияның тематикасын сипаттайтын кілт сөзден тұрады.

3.2.3 Терминдерді фильтрлеу

Ғылыми білім аймағындағы онтологияны құрастыру алгоритмінің келесі қадамы екі деңгейден тұратын $Terms_1$ терминдерінен алынған фильтрация болып табылады. Фильтрацияның бірінші деңгейінде терминнің өлшеміне жатпайтын кейбір жұптар жойылады. Ол үшін ары қарай тізбектелінген төрт өлшем қолданылады. $A \in Terms - A_1$ және A_2 екі сөзінен тұратын термин-кандидат болсын, сонда бұл өлшемдер мынандай жолмен қалыптасады:

Википедия онлайн энциклопедиясында A деп аталатын мақала бар;

- $\frac{hits("A\ is\ a\ term")}{hits(A)} > C_1$;
- $\frac{hits("A\ is\ a\ concept")}{hits(A)} > C_2$;
- $\frac{hits("A_1\ AND\ A_2")}{\min\,(hits(A_1),(hits(A_1))} > C_3$.

$hits(x)$ функциясы x сұранысына жауап ретінде Интернеттегі іздеу жүйесінен табылған парақшалардың санын білдіреді. Термин-кандидат аталған төрт өлшемнің біреуіне болса да сәйкес келетін фильтрдан өтіп кеткен бірінші деңгейі болып есептелінеді. $C_1, C_2, C_3 \in [0,1]$ сандары алгоритмнің параметрлері болып есептеледі.

Фильтрлеудің екінші деңгейінің мақсаты D білімінің берілген аймағымен байланысты болмайтын сөздердің жұбын жою болып табылады. Ол үшін мына өлшем қолданылады:

$$\frac{hits("A\ AND\ D")}{hits(A)} > C_4 \tag{3.18}$$

мұндағы A – термин-кандидат, D – білім аймағындағы берілген атауы, ал C_4 –алгоритмнің параметрі. Фильтрлеудің екі деңгейінен де сәтті өткен $Terms_1$ ішіндегі терминдердің жиынын $Terms_2$ арқылы белгілейміз. Ол $O_D: N_x^D = Terms_2$ онтологиясының және N_x^D экземпляр атауларының жиынын құрайды.

3.2.4 Ассоциативті қатынастарды анықтау

Келесі сатының мақсаты баланысты терминдердің жұптарын ерекшелеу болып табылады, яғни семантикалық жақын болып келетін жұптардың N_x^D жиынындағы мүмкін болатын термин жұптарымен байланысты. Екі терминнің арасындағы семантикалық жақындық деңгейін анықтау үшін Normalized Google Distance (NGD) кеңінен таралған шарты қолданады. A мен B – терминдер, ал N – ізденіс жүйесімен индекстелетін парақшалардың жалпы саны болсын. Сонда A мен B арасындағы NGD семантикалық жақындық деңгейі мына формула арқылы анықталады:

$$NGD(A,B) = \frac{\max\{\log hits(A), \log hits(B)\} - \log hits("A\ AND\ B")}{\log N - \min\{\log hits(A), \log(B)\}} \tag{3.19}$$

Осыдан кейін терминдердің барлық жұптарынан араларындағы бастапқы мәннен асатын жақындық деңгейі алынады:

$$Terms_s = \{(A,B) \in N_x^D * N_x^D | NGD(A,B) > C_{NGD}\} \tag{3.20}$$

C_{NGD} саны алгоритмнің параметрі болып табылады. A, B термин жұптары арасына $relatedTerm$ симметриялық қатынасы онтологияға $Terms_S$ жиынынан қосылады.

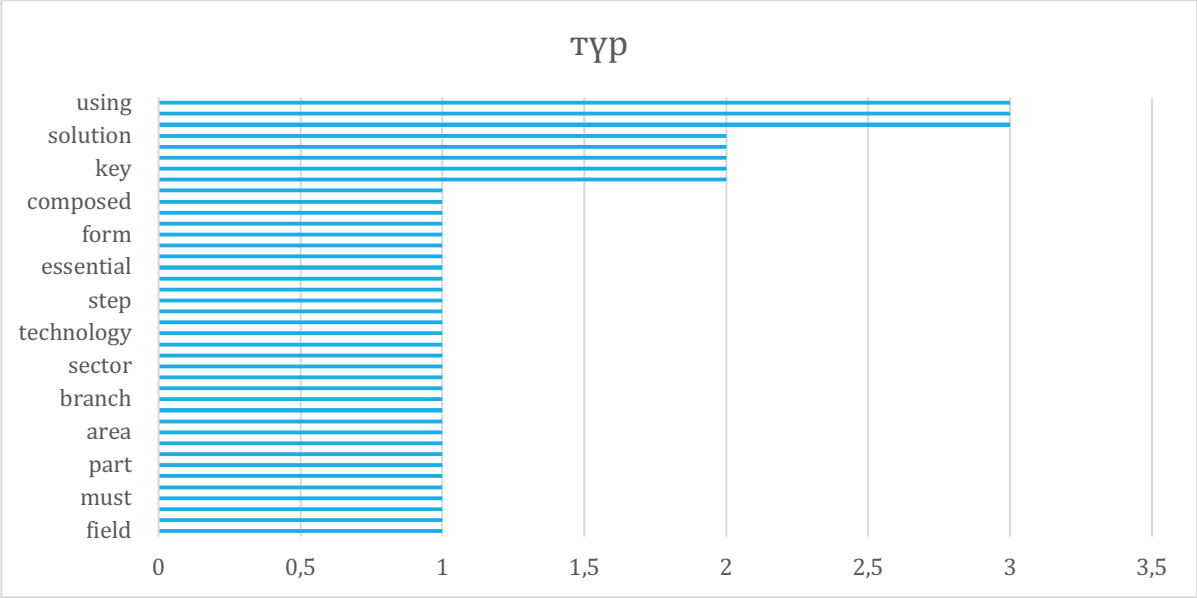
3.2.5 Иерархиялық қатынасты анықтау

Херст [42,с. 845] зерттеуін ғылыми бағыттағы терминдер иерархиясын құрастыру үшін турасынан қолдануға болмайды. Херст зерттеуі негіз болатын идеялардың қиындығына қарамастан оны қойылған мақсатқа жету барысында қолдануға болады.

Берілген зерттеу жұмысындағы зерттеуді жүргізу барысында ғылыми деңгейдегі иерархия түсінігін қалыптастыру үшін лингвистикалық шаблондар жасалынған. Негізгі шаблондар мынандай:

$$A\ is*\ keyword*prep(aux)?\ B \tag{3.21}$$

мұндағы A , B – иерархиялық байланыс ізденісі болатын терминдер, $keyword$ – ғылыми терминдер арасындағы $keywords$ байланысының құралған сөздігінен алынған кілт сөз, $prep$ – жалғаулар сөздігіндегі жалғаудың бір түрі, ал aux – көмекші сөздерге арналған сөздіктегі артикль немесе квантор. $keywords$ кілт сөздерінің сөздігі 40 сөзден тұрады, олар сурет 3.4 берілген. Іздеу жүйесі $Terms_S$ жиынына енген терминдер жұбы үшін ғана орындалады. Алдыңғы сатыда алынған семантикалық жақындық туралы деректерді қолдану ізденіс жүйесіне сұраныстар санын азайтуға мүмкіндік береді. Шаблон бойынша табылған сөз тіркестерінің мысалы ретінде “text categorization is a fundamental task in document processing” қолданылады. Мұндағы $A =$ “$text\ categorization$”, $B =$ “$document\ processing$”, $keyword\ =$ “$task$”, $prep\ =$ “in”, ал aux элементі қолданылмайды.



Сурет 3.4 - Терминдердің иерархиясын құрастыру кезінде қолданылған keywords кілт сөздерінің тізімі

$keywords$ сөздігінде көрсетілген сөздің негізгі формасынан басқа көпше түрдегі нұсқасы да қолданылады (сөзге “s”мен “es қосымшалары қолданылады). Кейбір кілттік сөздер байланысына келесі түрде берілген модификацияланған

шаблондар қолданылады (2 мен 3 типтері кесте 3.1 берілген):

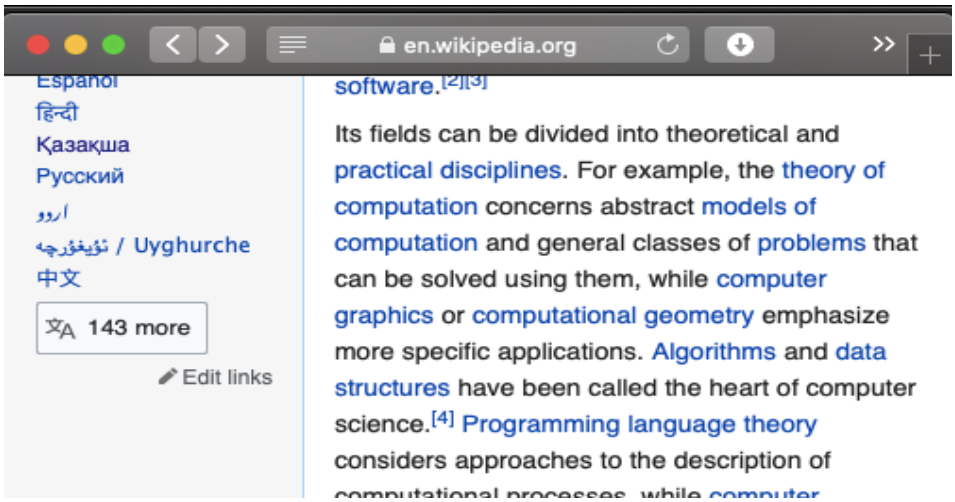
$$A\ is*\ keyword*prep_{ext}(aux)?\ B$$
$$A\ is*\ keyword*aux?\ B \quad (3.22)$$

мұндағы $prep_{ext}$ – “that” сөзінің көмегімен $prep$ қосымшаларының кең жиынында қолданылатын қосымша.

3.2.6 Терминдерді қазақ және орыс тілдеріне аудару

Жоғарғы сатыда сипатталынған ғылыми аймақ төңірегіндегі онтологияны құрастыру алгоритмі ағылшын тіліндегі онтологияны құрау үшін қолданылады. Қазақ және орыс тілдерінде сөйлем конструкциялары әртүрлі болғандықтан отандық ғалымдардың жасақтамаларына жүгінеміз, себебі бұл тапсырма диссертациялық жұмыстың негізгі тапсырмалар қатарына жатпайды.

Онтологияға қазақ және орыс тілдеріндегі терминдерді енгізу үшін келесі эвристикалық әдістемені қолдану талап етіледі. Оның негізгі идеясы Википедия энциклопедиясын құрайтын адамдардың қол еңбегін қолданады. Википедияда мақалалардың әр түрлі тілдерінде сілтемелері бар (сурет 3.5). Сол сілтемелер арқылы ағылшын тіліндегі терминдердің қазақ және орыс тілдерінде баламасы алынады.



Сурет 3.5 – Википедиядағы басқа тілдердегі мақалаларға сілтеме

N_x^D жиынындағы әрбір термин үшін дәл осындай атауы бар Википедиядағы мақалалардың бар екендігін автоматты түрде тексеріледі. Бұл мақаладағы қазақ және орыс тіліндегі нұсқасына сәйкес келетін негізгі термин сілтемене N_x^D онтологиясының жиынтығына қосылады.

3.2.7 Категорияларға жіктеу

Онтологияға жазылатын терминдерді жіктеу алгоритмі қолданылады. Термин бірнеше класта болуы мүмкін. Оларды келесі класқа жіктейміз кесте 3.1. Кестеде қазақ және орыс тіліндегі түсініктері берілмеген сурет 3.5.

Кесте 3.1 – Білім аймағындағы онтология түсінігінің тізімі

№	санат	№	санат	№	санат
1	field	11	algorithm	21	Paradigm
2	domain	12	concept	22	Technology
3	theory	13	Notion	23	Methodology
4	model	14	conception	24	Tool
5	problem	15	Idea	25	Practice
6	task	16	technique	26	Area
7	method	17	process	27	Phase
8	law	18	approach	28	Procedure
9	hypothesis	19	activity	29	Application
10	experiment	20	discipline	30	Mechanism

$A \in Terms_2$ терминіне жататын класстардың кішкентай жиынын анықтау үшін келесі формула бойынша $C \in N_C^D$ әрбір классына жататын оның деңгейі есептелінеді:

$$score(A, C) = \frac{hits(\text{“}A\ is\ a\ C\text{”})}{hits(A)} \tag{3.23}$$

Егер $score(A, C) > C_5$ онда A термині мен C классының арасында экземплярдың классқа жататындығын білдіретін $rdf:typ$ қатынас құрылады. C_5 саны алгоритмнің параметрі болып табылады. Тағы да ескере кететін жағдай, термин бірнеше класстарға бірдей уақытта ене береді.

Берілген алгоритмнің есептеу күрделілігін бағалауға мүмкіндік беретін теореманы құрастырамыз және дәлелдейміз.

3 теорема. R жиынынан бірнеше айдарға бөлінген Doc құжатының жиыны берілді делік. W - құжатта кездесетін сөздер жиыны және $m = |W|$ - сөздер саны болсын. $L(d)$ дегеніміз $d \in Doc$ – құжаттың ұзындығын көрсететін функция болсын (онда сөздердің саны жоқ). Құжат жиынындағы сөздердің саны $n = \sum_{d \in Doc} L(d)$ деп белгілейік. Онда алгоритм үшін келесі бағалау жолы дұрыс болмақ:

- алгоритмнің уақытша күрделілігі (нашар жағдайда)

$$O(min(m^4, n^2) + |R||Doc|min(m^2, n) + n);$$

- алгоритмнің кеңістік күрделілігі (нашар жағдайда)

$$O(min(m^4, n^2) + |Doc|min(m^2, n) + n).$$

Алгоритмнің уақытша күрделілігі арқылы тапсырмаларды орындауға қажетті қарапайым операциялардың максималды саны түсіндіріледі (арифметикалық, ізденіс жүйесіне бір рет қана жүгіну мен салыстыру операциясы). Кеңістік күрделілігі арқылы алгоритм жұмысы үшін ерекшелеп алуға қажетті жады ұяшықтарының максималды санын білетін боламыз.

Дәлелдеме. Реті бойынша ғылыми білім аймағындағы онтологияны құрастыру алгоритмінің әрбір сатысын қарастырамыз. Бірінші деңгейде онтология түсінігі атауларының жиынын құрастыруы жүзеге асырылады және

ол $O(1)$ тең болады.

Екінші деңгейде қолданылатын құжаттардың ішінен терминдерді белгілеу жүзеге асады. Ол үшін терминдерді ерекшелеу алгоритмі қолданылады. Оны анықтауға арналған бағалау түрі 2ші теоремадан алынған. Осындай жолмен, ғылыми білім аймағынындағы онтологияны құрастыру алгоритмінің екінші деңгейінің уақытша күрделілігі $O(min(m^4, n^2) + |R||Doc|min(m^2, n) + n)$, ал кеңістікке жататын – $O(min(m^4, n^2) + |Doc|min(m^2, n) + n)$ тең. Бұл деңгейде ізденіс жүйелерінің көмегіне жүгінуге болмайды. Қолданылатын дерек ретінде коллекция арқылы берілген құжаттар ғана қолданылады. Терминдерді ерекшелеу алгоритмнің жұмысының нәтижесінде алынған терминдердің санын t арқылы анықтаймыз.

Алгоритмнің үшінші қадамында алынған терминдік жұп сөздердің іріктелінуі іске асады. Сөздің әрбір жұбы үшін $A = (A_1, A_2)$ ізденіс жүйесіне қатысты 8 сұраныс орын алады: $hits(A)$, $hits(\text{“}A\ is\ a\ term\text{”})$, $hits(\text{“}A\ is\ a\ concept\text{”})$, $hits(\text{“}A1\ AND\ A2\text{”})$, $hits(A1)$, $hits(A2)$, $hits(\text{“}A\ AND\ D\text{”})$ және де Википедиядағы бір сұраныс. Соңғы формулада D құрастырылатын онтологияның білім аймағындағы берілген атауларды білдіреді. Бұл деңгей бойынша барлығы $8t$ сұраныс бойынша орындалады. Содан кейін алынған нәтиже бойынша салыстыру операциясы мен $O(t)$ арифметикалық операция іске асады. Осындай жолмен, бұл деңгейдің уақытша күрделілігі $O(t)$ тең. Бұл қадамның кеңістік күрделілігі $O(1)$ тең, себебі барлық операциялар реті бойынша және ешқандай терминге тәуелсіз күйде, ешқандай жаңа ақпаратты есте сақтамай-ақ іске асады.

Төртінші деңгейде ассоциативті қатынас орындалады. Әрбір терминдік жұп NGD функциясымен есептелінеді. Әрбір A, B терминнің жұбы үшін біршама уақыт қажет. Сол кезде ізденіс жүйесіне сұраныс тасталынады: $hits(\text{“}A\ AND\ B\text{”})$. Осыдан кейін бастапқы мәні берілген NGD мәні жұптан алынып тасталынады. Осындай жолмен, деңгейдің уақытша күрделілігі $O(t^2)$ тең. Бұл деңгей бойынша $\frac{t(t\ 1)}{2}$ аспайтын ізденіс жүйесіндегі сұраныстан тұрады. Деңгейдің кеңістік күрделілігі $O(t^2)$ тең, себебі терминдердің статистикалық жақын жұптарын сақтау қажет.

Алгоритмнің бесінші сатысында терминдер иерархиясы құрастырылады. Араларында алдыңғы деңгейде анықталған байланысы бар терминдердің әрбір жұптары үшін (A, B) “$A\ AND\ B$” түріндегі ізденіс жүйесіндегі сұраныс орындалады. Атап өтейік, $hits$ түріндегі сұраныспен салыстырғанда бұл сұранысқа жауап ретінде сұраныс сөзі кездесетін веб-парақшалардың саны ғана емес, сонымен қатар сұраныстағы сөз кездесетін қысқаша мәтіндер де алынады. Бұл сатының уақытша күрделілігі, кеңістік күрделілігі $O(t^2)$ тең. Мұнда терминдердің әрбір жұбы үшін иерархиялық байланыс ақпаратты сақтау қажет.

Алтыншы сатыда Википедиядағы ақпараттың көмегі арқылы ағылшын терминдерінің қазақша және орысша эквиваленттерін іздеу жұмысы жүргізіледі. Әрбір термин үшін Википедияға бір рет қана сұраныс жіберіледі. Операцияның жалпы саны $O(t)$ тең. Кеңістік күрделілігі $O(t)$ тең, себебі әрбір термин үшін максимум ең аз дегенде бір эквивалентті сақтап қалу маңызды.

Алгоритмнің ең соңғы жетінші деңгейінде N_C жиынындағы категорияларға терминдер бөлінетін болады. Бұл жиын қарастырылып отырған ғылыми бөлімге байланысты емес, және де ол реттілік бойынша 30 сөзден тұрады. Әрбір термин үшін іздеу жүйесінде $|N_C|$ сұраныс орын алады бұл деңгейдің уақытша күрделілігі $O(t)$ тең. Кеңістік күрделілігі де $O(t)$ тең, себебі әрбір термин үшін ең аз дегенде $|N_C|$ байланысты сақтап қалу қажет.

Әрбір сатыға арналған күрделілік бағасын біріктіру арқылы алған мөлшерді қорытынды көрсеткіш ретінде шығарамыз. 2ші теорема арқылы дәлелденген терминдерді ерекшелеп алу алгоритмінің көмегімен t шығарылған терминдерінің саны үшін бағаны қолданатын боламыз. Осындай жолмен, ғылыми білім деңгейіндегі онтологияны құрастыру алгоритмінің уақыт бойынша күрделілігі келесідей болмақ:

$$O(1) + O(|R||Doc|\min(m^2, n) + n) + O(t) + O(t^2) + O(t^2) + O(t) + O(t) = O(|R||Doc| \min(m^2, n) + n + t^2) = O(\min(m^4, n^2) + |R||Doc|min(m^2, n) + n)$$

Алгоритмнің кеңістік күрделілігі мынаған тең:

$$O(1) + O(|Doc| min(m^2, n)) + O(1) + O(t^2) + O(t^2) + O(t) + O(t) = O(|Doc| min(m^2, n) + t^2) = O(min(m^4, n^2) + |Doc| min(m^2, n)).$$

3 бөлім бойынша тұжырым

Берілген бөлімде ғылыми ақпаратты басқару жүйесінде қолданылатын екі кілттік алгоритмдердің сипаттамасы бар. Мәтіндер коллекциясынан терминдерді ерекшелеп алу алгоритмі статистикалық болып саналады және төрт қадамнан тұрады. Ғылыми білім аймағындағы онтологияны құрастыру алгоритмі ғылыми конференциялар анонсы терминді ерекшелеу үшін қолданылатын дерек ретінде қолданды. Терминдердің арасындағы қатынасты анықтау үшін Интернеттегі ізденіс жүйесі арқылы табылған ақпарат қолданылды. Алгоритм семантикалық әдіс негізінде жасалды.

Әзірленген алгоритмдердің сипаттамасы осы жұмыстың ғылыми ақпараттық қорытындысы мен есептеу жүйесіне сәйкес үлгі құрайды. Бұл алгоритмдер ғылыми білім аймағында отнологияны құрап және терминдерді ерекшелеп алу мүмкіндігі туындайтын басқа да жүйелерде қолданылуы мүмкін.

4 ЗЕРТТЕУ НӘТИЖЕЛЕРІН БАҒАЛАУ ЖӘНЕ АПРОБАЦИЯЛАУ

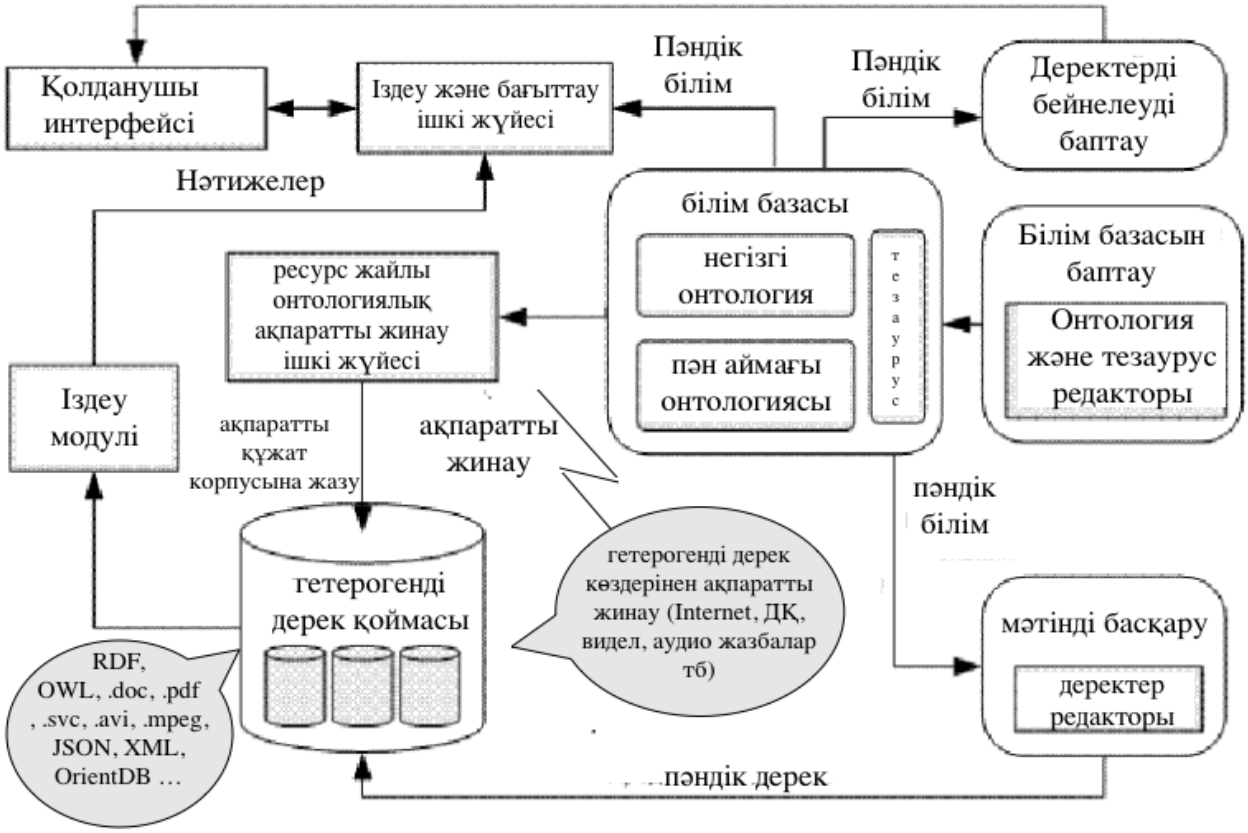
Бұл бөлімде ұсынылған үлгі, әдіс және алгоритмді жүзеге асыру мақсатында интелектуалды жүйе прототипі құрастырылды.

Екі онтологияның арасындағы мағыналық байланыстарды автоматты түрде анықтауға және орнатуға, сондай-ақ бір жүйеден екінші жүйеге ақпарат қосу сұраныстарын құруға мүмкіндік береді. Онтологиялық ұғымдардың байланысы негізінде объектілік схемалардың сәйкес элементтері біріктіріледі.

Бағдарламалық қамтамасыз ету жүйесінің тұжырымдамалық үлгі жасалды. Бағдарламалық жасақтама жүйесінің жұмысының нәтижелері көрсетілген. Оның тиімділігі мен пайдалылығын бағалау үшін тәжірибе жасалды.

4.1 Автоматтандырылған біріктіру жүйесін құру

Диссертациялық жұмыста сипатталған әдістер мен алгоритмдерді жүзеге асыру үшін гетерогенді дерек көздерден қажетті ақпаратты шығарып, олармен байланысты деректер түрінде ұсынуға арналған бағдарламалық жасақтама құрылды. Жүйе таратылған, біріктірілген жүйелерде сақталған ақпаратты онтологиялық үлгіге дербес көрсетуі керек. Сонымен қатар, алынған үлгіге қатысты деректер ретінде қол жетімділікке арналған интерфейсті қамтамасыз етуі керек, осылайша сурет 4.1 көрсетілген бірнеше жұмыстарда сипатталған деректерге қол жетімділікке арналған онтологиялық тәсіл жүзеге асады [43-49].



Сурет 4.1 – Жүйелік архитектура

Бұл тарау нәтижесінде пайда болған жүйені, оның архитектурасын қарастырады, сонымен қатар аналогтармен: деректерді біріктірудің дәстүрлі жүйелері мен семантикалық жүйелер салыстырады.

4.2 Жүйеге қойылатын талаптар

Жүйені жобалау үшін, ең алдымен, жүйеге сәйкес болатын талаптардың тізімін жасау қажет болды. Талаптар мақсатқа сәйкес деректерді біріктірудің семантикалық жүйелері талданды.

- Функционалды талаптар;
- Стандарттардың талаптары;
- Қауіпсіздік талаптары;
- Өнімділікке қойылатын талаптар.

4.2.1 Функционалды талаптар

Сипатталған жүйелер мен қадамдарды талдау негізінде жасалған әдіс пен алгоритмдер, жүйеде міндетті түрде қолданушының негізгі сценарийлері анықталды.

Пайдаланушы сценарийі 1. Жүйені қолдана отырып, пайдаланушы желі ішіндегі кез-келген реляциялық дерекқорының құрылымы туралы ақпаратты онтология түрінде RDF форматында ала білуі керек.

Ақпаратты жеке дерекқор үшін де, олардың барлық мағыналары үшін олардың семантикалық байланыстарын ескере отырып бірге беруге болады.

Пайдаланушы сценарий 2. Жүйені қолдана отырып, пайдаланушы мәліметтер қорында сақталған ресурстар туралы семантикалық мета деректерді ала алады. Мұндай деректерге қол жеткізу үшін SPARQL сұрауларын қолданады, оларды желі ішінде орналасқан қолданбалы серверге бағыттайды.

Бұл сценарийлер негізгі болып табылады және дамыған жүйенің барлық мүмкіндіктерін сипаттамайды. Пайдаланушы жүйемен ыңғайлы жұмыс жасау үшін, көрсетілген сценарийлер шеңберінде бағдарламаның белгілі бір функционалдық талаптарына сай болуы қажет.

ФТ-1 талабы. Жүйе реляциялық дерекқорының барлық кең таралған түрлерін қолдауы және кеңейту мүмкіндігі болуы керек. Oracle, MSSQL Server, PostgreSQL, MySQL сияқты ДҚБЖ қолдау көрсетілуі керек. Әр түрлі МҚБЖ қолданатын жүйелерді бір уақытта біріктіруге болады.

ФТ-2 талабы. Жасалып жатқан жүйе біріктірілген дерекқордың әрқайсысының құрылымы туралы ақпарат алып, оны RDF/OWL форматында ұсынуы керек.

ФТ-3 талабы. Дерекқордың құрылымы туралы ақпаратты қамтитын бөлек мета үлгілерді бірыңғай онтологияға автоматты түрде біріктіру қажет. Біріктіруде ақпараттық жүйелер арасындағы мағыналық байланыстар ескерілуі керек.

ФТ-4 талабы. Жүйе құрылымның жалпы онтологиясына негізделген ақпараттық ресурстар қорында сақталған семантикалық мета-сипаттамалар шығаруы керек.

ФТ-5 талабы. Ресурстық мета деректерді шығару кезінде қайталанатын

ақпаратты автоматты түрде анықтау қажет.

ФТ-6 талабы. Әзірленген жүйе желі ішіндегі SPARQL сұраулары арқылы семантикалық мета деректерге қол жеткізу мүмкіндігін қамтамасыз етуі керек.

ФТ-7 талабы. Барлық қажетті жүйелік параметрлер мен біріктірілген ақпараттық жүйелер конфигурациясы қолданушының веб-интерфейсі арқылы жасалуы керек.

ФТ-8 талабы. Жүйе пайдаланушыға веб-интерфейс арқылы немесе бағдарламалық қамтамасыз ету API арқылы өзара әрекеттесу әдісін таңдауға мүмкіндік беруі керек.

ФТ-9 талабы. Біріктірілген жүйелерден мета сипаттамаларды алу үрдісі кестеге сәйкес, конфигурацияланатын уақыт аралығында жүзеге асырылуы керек. Осылайша, тұрақты сұраныстардың арқасында біріктірілген дерекқордың жүктемені одан әрі арттырмай, онтологиялық үлгіге заманауи мета деректердің болуына кепілдік беруге болады.

Бұл талаптарда әзірленген әдісті ескере отырып тұжырымдалған және ұқсас мақсаттағы талданатын жүйелерде қолданылатын ең жақсы шешімдерді қамтылды.

4.2.2 Стандарттық талаптар

Басқа мағыналық жүйелермен өзара әрекеттесу үшін стандартталған технологиялар мен әдістерді қолданған жөн. Осыған байланысты стандарттарға қойылатын талаптарды ерекше атап көрсетілді.

СТ-1 талабы. Жүйе w3c консорциумы ұсынған стандартты Semantic Web технологияларын қолдануы керек.

Онтологиялық үлгі OWL тілін пайдаланып RDF форматында сипатталуы керек. Деректерге SPARQL сұраныс тілі арқылы қол жеткізіледі. Жүйе қолданыстағы жоғары деңгейлі онтологияларды атау кеңістігі механизмдері арқылы пайдалануды қолдауы керек.

СТ-2 талабы. Жүйе реляциялық дерекқорын салыстыру стандарттары шеңберінде ұсынылған шешімдерді жүзеге асыруы керек.

СТ-3 талабы. Біріктірілген жүйелерге SOAP веб-қызметтері және JDBC дерекқорының адаптері сияқты стандартталған құралдардың көмегімен қол жеткізілу керек.

4.2.3 Қауіпсіздік талаптары

Жүйеде қауіпсіздік талаптарына жауап беретін бірқатар талаптары бар.

ҚТ-1 талабы. Жүйеде пайдаланушыларды рөлдерге бөліп, авторизация және аутентификация механизмдері қосу керек. Әр пайдаланушының жеке кабинеті және белгілі бір рөлі: пайдаланушы, сарапшы, әкімші болуы керек. Желідегі LDAP хаттамасын қолдау қажет.

ҚТ-2 талабы. Біріктірілген ақпараттық жүйелер арасында ақпарат беру, сондай-ақ веб-интерфейспен жұмыс SSL хаттамасының көмегімен шифрланған канал арқылы жүзеге асырылуы керек.

4.2.4 Өнімділікке қойылатын талаптар

Бұл талаптар жүйенің өз функцияларын қаншалықты тез және тиімді орындауы керектігін анықтайды. Жылдамдық, өткізу қабілеттілігі, бірнеше пайдаланушы бөліскен кезде жүктеу және т.с.с. параметрлерді анықтау керек. Жүйенің өнімділігі кіріс деректерінің көлеміне байланысты, бірақ ол төменде көрсетілген негізгі талаптарға сай болуы керек.

ӨТ-1 талабы. Жүйе бір уақытта кем дегенде 20 пайдаланушыға қызмет көрсете алуы керек. Пайдаланушылар санының серверлер мен жүктемені үлестіруді көбейту есебінен арттыруға болады.

ӨТ-2 талабы. SPARQL API қалыпты жүктеме кезінде 2 секунд ішінде және жоғарылаған жүктеме кезінде 5 секунд ішінде жауап беруі керек.

ӨТ-3 талабы. Мета деректерді шығару тікелей шығыс деректердің көлеміне байланысты, бірақ бір жазбаны өңдеуге кететін уақыт 0,5 секундтан аспауы керек.

Жүзеге асырылған жүйені, егер ол жоғарыда сипатталған талаптарға сәйкес келсе, жұмыс істей алады деп санауға болады. Жүйені жобалау кезінде олар ескеріліп, енгізілді және сәйкестік матрицасы құрылды. Осы талаптардың негізінде өнімнің сапасын тексеру үшін тестілік сценарийлер жазылып, жүйе тексерілді.

4.3 Функционалды үлгі жасау

Ұсынылған функционалдық талаптарға сәйкес ақпараттық ресурстарды біріктірудің автоматтандырылған жүйесінің функционалдық үлгісі жасалды. Үлгі жасау үшін IDEF0 стандартындағы SADT әдіснамасы қолданылды. Бұл кез-келген пәндік аймақ объектісінің функционалды үлгісін құруға арналған ережелер мен үрдістер жиынтығы. Әдіске сәйкес барлық функционалды компоненттер блоктар мен доғалардан тұратын диаграмма түрінде бейнеленуі керек. Маңызды ерекшелігі - үлгі бөлшектерінің деңгейінің біртіндеп жоғарыланады. Сурет 4.2 ең аз бөлшектермен А-0 диаграммасы көрсетілген.

Сурет 4.2 - Функционалды үлгі, А-0 диаграммасы

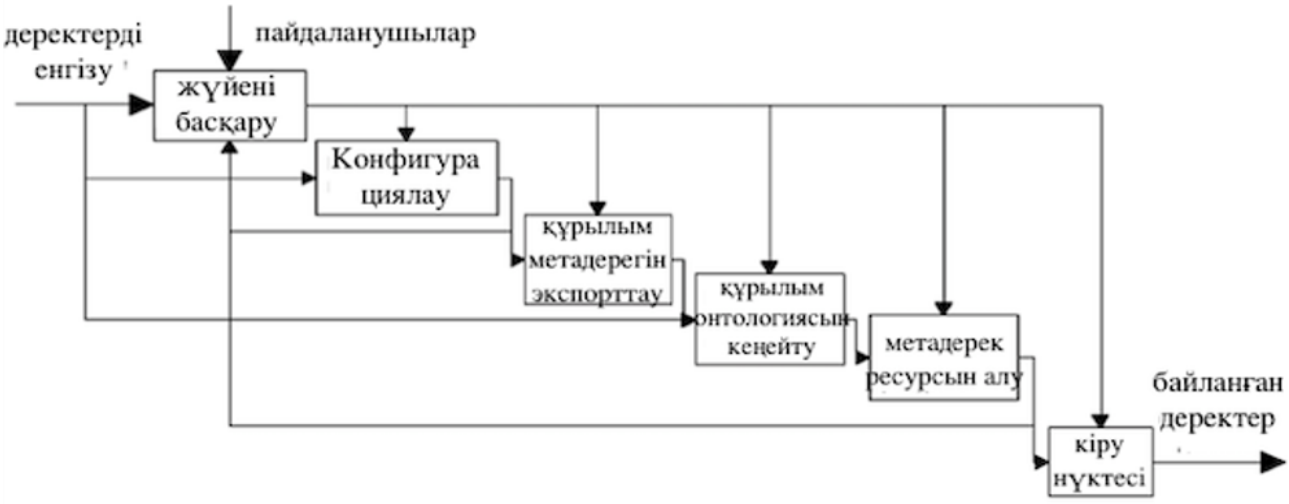
Жүйе әртүрлі кіріс деректерін алады: біріктірілген жүйелер туралы ақпарат, пайдаланушылар туралы ақпарат, басқару командалары. Жоғарыда блок басқару пәрмендерін төменнен жүйеге кіретін үлгілерден алады. Шығарылған кезде біріктірілген ақпараттық жүйе мета деректерді қамтитын онтология құрады.

Детализация функционалды үлгіні ағаш түрінде ұсынылған, мұнда әр түйін үрдіс болып табылады:

Автоматтандырылған жүйе

- ▫ Жүйені басқару
- Пайдаланушы құқықтары мен рөлдерін белгілеу
- Жүйені конфигурациялау және басқару
- ▫ Біріктірілген ақпараттық жүйені конфигурациялау
- Өзара әрекеттесу тәсілін белгілеу
- Байланыстарды орнату
- ▫ Құрылым мета деректерін экспорттау
- Ақпараттық жүйеге қосылу
- Құрылымдық ақпарат алу
- Ұқсас құрылымдық элементтерді талдау
- Талдау негізінде семантикалық сілтемелер қосу
- Онтологияны экспорттау
- ▫ Онтология құрылымының кеңеюі
- Жоғары деңгейлі онтологияларды қосу
- Домендік онтологияларды қосу
- Семантикалық байланыстарды қосу
- Кеңейтілген онтологияны экспорттау
- ▫ Ақпараттық ресурстардан мета деректерді алу
- Кеңейтілген құрылымдық онтологияның импорты
- Ақпараттық жүйеге қосылу
- Мета деректерді шығару
- Ұқсастыққа талдау
- Талдау негізінде мағыналық сілтемелер қосу
- Онтология құрылымына негізделген семантикалық сілтемелерді қосу
- Онтологияны репозитарийге экспорттау
- ▫ Кіру нүктесін ұсыну
- Репозиторийден мета деректерді алу
- Кіру нүктесінің серверін бастау

SADT әдісі бойынша бірінші деңгейдің функционалды үлгісінің сызбасы сурет 4.3 көрсетілген.



Сурет 4.3 - А-1 Детальданған диаграммасы

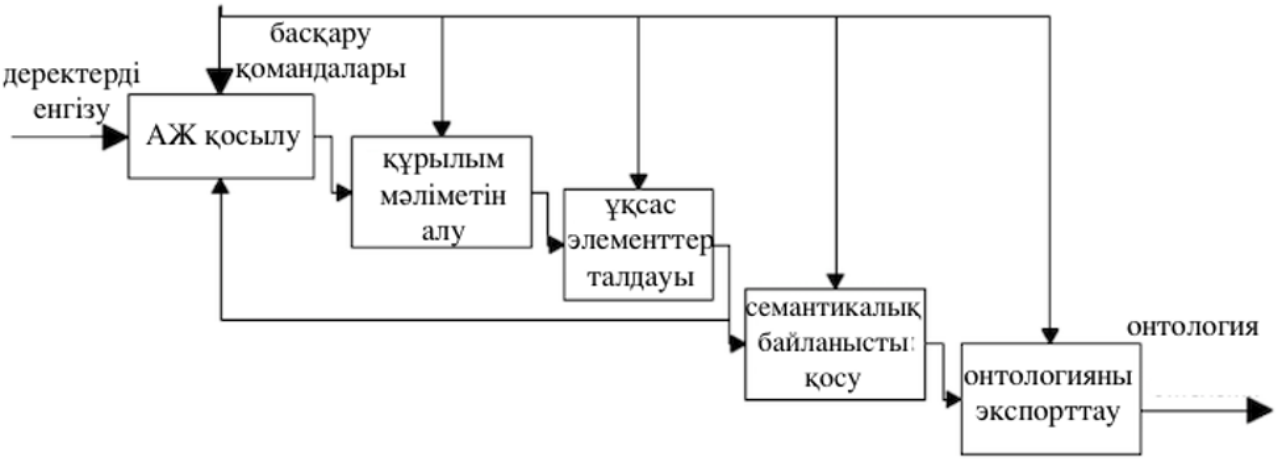
Жүйе кіріс деректері мен пайдаланушының командаларын алады. Жұмыс кезінде онтологиялық құрылымды қалыптастыру үшін біріктірілген жүйелер базасынан әртүрлі ақпарат экспортталады. Семантикалық байланыстар санын арттыру үшін онтология жұмыс барысында бірнеше рет экспортталады және түрлендіріледі.

Шығару кезінде алынған мета деректер мен онтология кіру нүктесінің көмегімен байланысқан деректер ретінде ұсынылады.

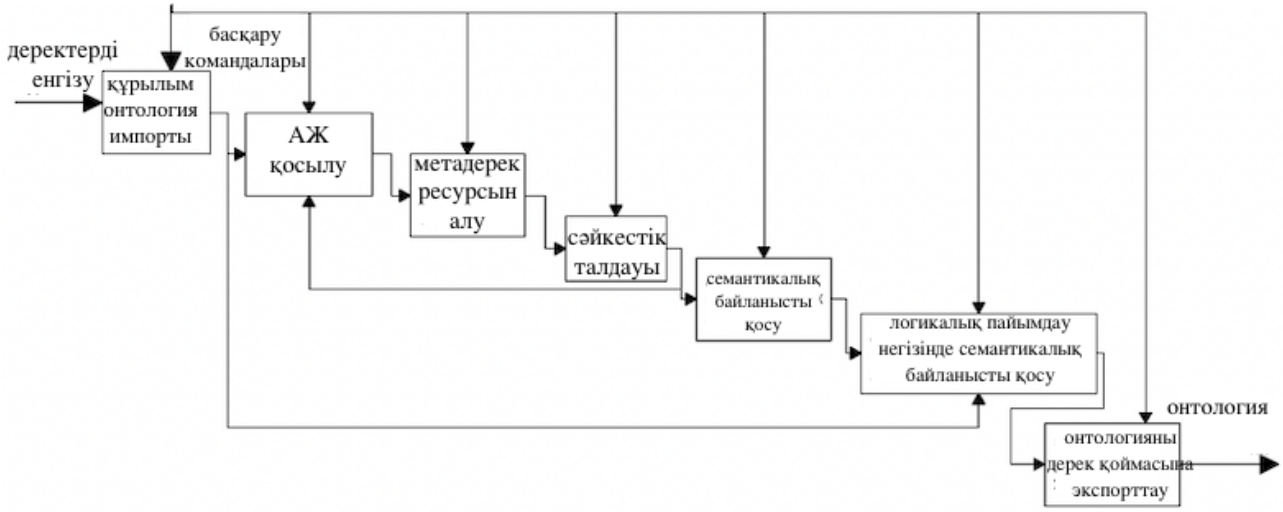
Сурет 4.4 құрылымның мета деректерін экспорттау үрдісінің детализациясы көрсетілген.

Бұл кезеңде біріктірілген ақпараттық жүйелердің құрылымы алынады. Кіріске жүйелер мен басқару командалары туралы мәліметтер беріледі. Құрылымдық ақпаратты шығару автоматты түрде жүреді және құрылым онтологиясы шыққан кезде жасалады.

Ақпараттық ресурстардың мета деректерін шығару үрдісінің детализациясы сурет 4.5 көрсетілген.



Сурет 4.4 – Құрылым мета деректерін экспорттау, А-1.3 диаграммасы



Сурет 4.5 – Ресурстардың мета деректерін экспорттау, А-1.5 диаграммасы

Бұл үрдіс алдыңғы үрдіске ұқсас, алайда ол ақпараттық ресурстардың мета деректерін шығаруға бағытталған. Маңызды айырмашылық - онтология құрылымына негізделген мета деректер арасында қосымша байланыстар орнату үшін логикалық пайымдау механизмін қолдану [48,с. 50]. Кеңейтілген құрылымдық онтология және басқару командалары кіріске жіберіліп онтология құрылады, ол ақпараттық жүйе құрылымы туралы ақпаратты да, оларда сақталған ресурстардың мета деректерін де қамтиды.

Осылайша, жұмыс барысында жүйеде үрдістер мен есептеулерді сипаттайтын функционалды үлгісі құрылды.

4.4 Терминдерді ерекшелеу алгоритмінің пайдасын зерттеу және бағдарламалық іске асыру

4.4.1 Тиімділікті бағалау әдістемесі

Терминдерді ерекшелеу алгоритмі JAVA тілінде іске асырылған [50]. Көптеген әдістер бір ғана сөзді ерекшелеуге мүмкіндік беретін, немесе айдарларға (бөлімдер, егер ол бар болса, қолданылмайды) бөлінбеген құжаттардан терминдерді ерекшелеуге бағытталған. Аталған жағдайға байланысты толықтай онтологиялық әдістермен терминдерді ерекшелеу алгоритміне тура салыстыру жұмысын жүргізу мүмкін емес.

Тестілеу барысында құжаттарды ұсынудың векторлы үлгісі қолданылды. Терминдерді ерекшелеу алгоритмі тапсырманың мөлшерін төмендете отырып, маңызы бар сөздерден қысқартылған базис орнатылды. Шығыс құжаттары мен олардың айдарларға бөлінуі туралы ақпараттарды сақтау маңызды шарттардың бірі. Дәл осы шарт құжаттарды ұсынудың векторлы үлгісінің шеңберінде шешілетін кеңінен таралған екі тапсырманың көмегімен формалды түрде тексерілуі мүмкін. Бұл тапсырмалар деректерді өңдеудегі көптеген заманауи жүйелерде кілттік роль атқарады.

Базистегі ақпаратты сақтаудың шартын тексеру үшін айдарларға бөлінген екі құжат алынды. Таңдап алынған құжаттың бірі үйретуші ретінде

қолданылады, ал екіншісі – тестік болады. Олар құжаттардың санына тең болады, екі таңдауда да айдарлар біркелкі болады.

Үйретушіні таңдаудың көмегімен жүйені оқыту жұмысы жүргізіледі. Классификацияның көрсеткіші бойынша бірнеше алгоритм айдарларға шығу жолымен бөліну қалпын сақтайды, ал кластеризация – кластерлердің қаншалықты шағын болып келетіндігіне байланысты.

Базистегі ақпаратты сақаудың шартын тексеру үшін мынандай әрекеттер орындалады. Айдарларға бөлінген екі құжаттар сұрыптамасы алынады. Таңдап алынғандардың бірі үйретуші, ал екіншісі тест ретінде қолданылады. Екі таңдауда да айдарлар біркелкі және құжаттар саны тең болады. Үйретуші сапасында таңдалған сұрыптамада жүйені оқыту жүзеге асырылады, демек, алдын-ала белгіленген мөлшердің қысқартылған базисін көрсететін терминдер ерекшеленеді. Содан кейін білім беруші және тесттік таңдау құжаты арқылы құжатқа і-ші номерімен термин-жұптардың енетін мөлшері бар і-ші координатасы бар, осы базистің негізінде кеңістікте нүктелер қойылады. Осындай жолмен, N құжатынан бір жағынан кластерлерге – шығыс айдарының көрінісі, және екінші жағынан екіге – үйретуші және тесттік таңдау (әрбір жиын түрі сәйкес келетін таңдаулардан алынған құжаттардан құралған), бөлінген N нүктелері туындайды. Кейбір айдардарға сәйкес келетін кластер осы айдардың құжаттарының көріністік нүктесі арқылы құралған. Алгоритмнің тиімділігінің нәтижесі классификация мен кластеризация әдістерінің көмегімен кластерлер конфигурациясын тестілеуден тұрады. Классификацияның көрсеткіші бойынша бірнеше алгоритм айдарларға шығу жолымен бөліну қалпы сақталды.

Классификация алгоритмінің көмегімен жасалатын бағалау жолы. Классификацияның көмегі арқылы бағалау келесі көрсетілген жол арқылы орын алады. «Жанындағы көршілер» әдісінің көмегі арқылы мәтіндік таңдау нүктесінің классификациясы орындалады, ол үшін үйретушіні таңдау нүктесінің кластерін бөлу әрекеті қолданылады. Жіктелудің нәтижесінде құжатты айдарлардың біріне жеткізетін тесттік таңдаудың әрбір нүктесі кластерлердің біріне келіп түседі – яғни, шығыс айдарының көрінісіне. Сонымен қатар нүкте айдарлардың біріне басынан бастап қатысты болып келетін құжатты ұсынады. Осындай жолмен, нүктелерді жіктеудің нәтижесінде тесттік таңдау құжатының нүктелерінің жіктелуі тағы бір рет айдарларға бөлінеді. Ары қарай бұл таралымның айдарларға бөлінетін құжаттармен қаншалықты сәйкес екендігі анықталады. Ол үшін жіктеуден кейін таңдаудың жалпы құжат санына басынан бастап тиісті болған айдарға келіп түскен тесттік таңдау құжаттарының санына деген пайыздық қатынас анықталады. Алынған сан қаншалықты көп болса (0% бастап 100% дейінгі аралықта), соншалықты алгоритмді қарастырудың тиімділігі артады.

Кластеризация алгоритмі арқылы бағалау. Кластеризация әдісі арқылы бағалау тесттік таңдау нүктелері арқылы жасалынған кластерлерге нүктелерді кластерлеуді бағалаудың стандартты әдістерінің бірі қолданылды.

4.4.2 Тестілеудің нәтижесі

Алгоритмдердің тиімділігін салыстыру мақсатында 5 мың құжаттың ішінен 1,1 миллион сөзден тұратын таңдаудан 21 мың сөзді құрады. Үйретушіні таңдау 3591 құжатты құрады, ал қалған құжаттар тесттік таңдауды құрады.

Терминдерді ерекшелеу алгоритмі әрбір сынға бір-бірден төрт параметр қолданылады. Параметрлердің тиімді мәнін анықтау үшін әртүрлі жиынтықтардың тестіленуі жүргізілген. Нәтижесінде терминдерді ерекшелеу алгоритмі максималды тиімділікті құрайтын келесі мәнге ие болады:

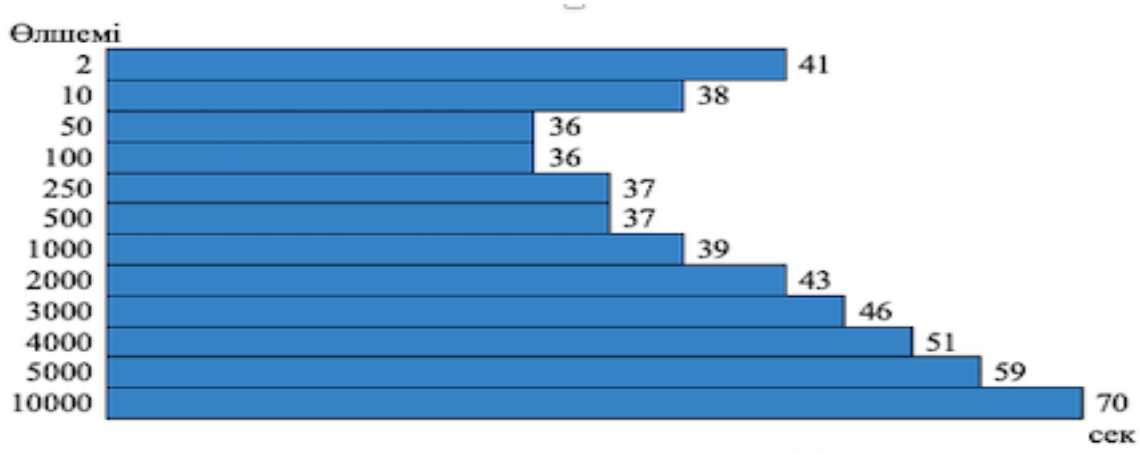
- MAX_DIST– 1 (сөздер қасында тұруы керек) – жұбын құрайтын сөздердің арасындағы максималды қашықтық;

- MIN_FREQ – M2 – 5 көпшесіндегі іріктеу үшін жиынтыққа енетін жұптардың минималды мөлшері;

- MIN_DISCR – M3 – 0.7 көпшесіндегі іріктеу үшін тән қызметтің минималды мәні;

- MAX_FREQ_RATIO – мәні бар айдарлардың өлшеміндегі константтың максималды мәні – 0 (фильтр өшірулі). Бұл параметрдің мәні мынада: MAX_FREQ_RATIO реттіктен аспайтын жұптың өлшемінен асатын жұптардың екуінің де сөздерінің санында айдардың жұбы үшін белгілі бір мән болуы керек. Тестілеуден күтпеген жағдай мәні бар айдарлардың өлшемдері өшірілген кезде жіктеудегі нақтылы мән арта түсетіндігі болды. Бұл қорытындыны соңғы нәтиже емес, себебі басқа үйретушіні таңдауда параметрлердің мәні басқаша болады.

2ші теорема бойынша алгоритмінің уақытша күрделілігі $O(||R|||Doc||min(m2,n)+n)$ ретінде бағаланады, мұндағы n– жиынтықтағы құжаттардың ішіндегі барлық сөздің жалпы саны, m – әртүрлі сөздердің саны, |Doc| – құжаттардың саны, а |R| – айдарлар саны. TF-IDF алгоритмінің уақытша күрделілігі O(||Doc||m) сурет 4.6 құрайды.



Сурет 4.6 – Терминдерді ерекшелеу алгоритмінің жұмыс уақыты

LSI, TF-IDF және терминдерді ерекшелеу алгоритмінің салыстырмалы жұмыс уақыты сурет 4.7 көрсетілген. Әрбір алгоритм жіктелу нақтылығының максималды нәтижесіне жететін тиімді мәні бар толықтай кеңістіктің өлшемдік параметрімен есептелді. Алгоритм үшін терминдерді ерекшелеу мәні 10000 тең,

LSI алгоритмі үшін – 250. TF-IDF алгоритмі толықтай кеңістіктің өлшемін таңдап алуға мүмкіндік бермейді. Ол жиынтықтағы ерекше сөздердің жалпы санына тең, яғни бұл жағдайда 21000 тең.

Сурет 4.7 – LSI, TF-IDF және терминдерді ерекшелеу алгоритмінің салыстырмалы жұмыс уақыты

Жадымен жұмыс істеуді ұйымдастыру. Мәтінді өңдеуге байланысты тапсырмалардың ерекшеліктерінің бірі үлкен деректермен жұмыс істеу болып табылады. Соған байланысты бағдарламадағы жадымен жұмыс істеуді ұйымдастыруға сай келетін бірнеше ерекшеліктерді атап өтейік. Алгоритмнің сипаттамасынан алгоритмді іске асырушы нұсқаны алайық: M_0 жиынын құрау, содан кейін M_1 кірмеген жұпты одан алып тастаймыз, осылай M_4 қорытынды жиыны шыққанға дейін жалғастыра береміз. Алайда, бұл жағдайда бағдарлама шамадан тыс көп оперативті жадыны талап етер еді (миллион сөзге таңдау жасау үшін бірнеше гигабайт).

Жадымен жұмыс істеуді ұйымдастырудың басқа нұсқасы 2 теореманы дәлелдеу барысында қолданылған идеяны ұстану. Алгоритм итерационды түрде іске асырылған: *PW* жұбының шығыс жиыны қуатқа тең (айтарлықтай көп емес) ұсақ жиындарға бөлінеді, содан кейін алгоритм әрбір ұсақ жиындарға ретімен қолданылды (яғни одан шығатын мыналар: M_1, M_2, M_3, M_4) сонымен қатар жадыда M_4 кірген жұптар ғана қалды. Алгоритмнің итерационды түрде іске асырылуы компьютердің жадысына қойылатын талаптарды айтарлықтай азайтуға мүмкіндік берді.

Терминдерді ерекшелеу алгоритмі үшін өлшем функциясын тестілеу. Терминдерді ерекшелеу алгоритмінде қолданылатын айдардағы жұптың өлшемінің функциясы үшін f мен g функцияларына үміткер ретінде келесі функциялар таңдалынды (мұндағы, алдыңғы жолғы сияқты, r айдары $d_1, \ldots d_k, k > 1$ құжаттарынан тұрады):

1. $\frac{\sqrt{x}}{\ln(1+y)}$ (графикте 1 санымен белгіленген) (4.1)
2. $\frac{x^2}{y}$ (4.2)
3. $\frac{x\sqrt{x}}{y}$ (4.3)
4. $\frac{x(1+\ln(1+x))}{y}$ (4.4)
5. $\frac{xln(1+x)}{y}$ (4.5)

6. $\frac{x}{\ln(1+y)}$ (4.6)

7. $\frac{x}{y}$ (4.7)

- $g(x_1, \ldots, x_k)$:

1. $\sum_{i=1}^{k} \ln(1 + x_i)$ (графикте $g1$ деп белгіленген);
2. $\prod_{i=1}^{k} (1 + x_i)$ $(g2)$;
3. $(\sum_{i=1}^{k} x_i) \cdot |\{x_i | x_i > 0\}| (g3)$

Көрсету үшін аргументтердің аз мөлшеріне тәуелді болып келетін функцияның қарапайым түрлері алынған:

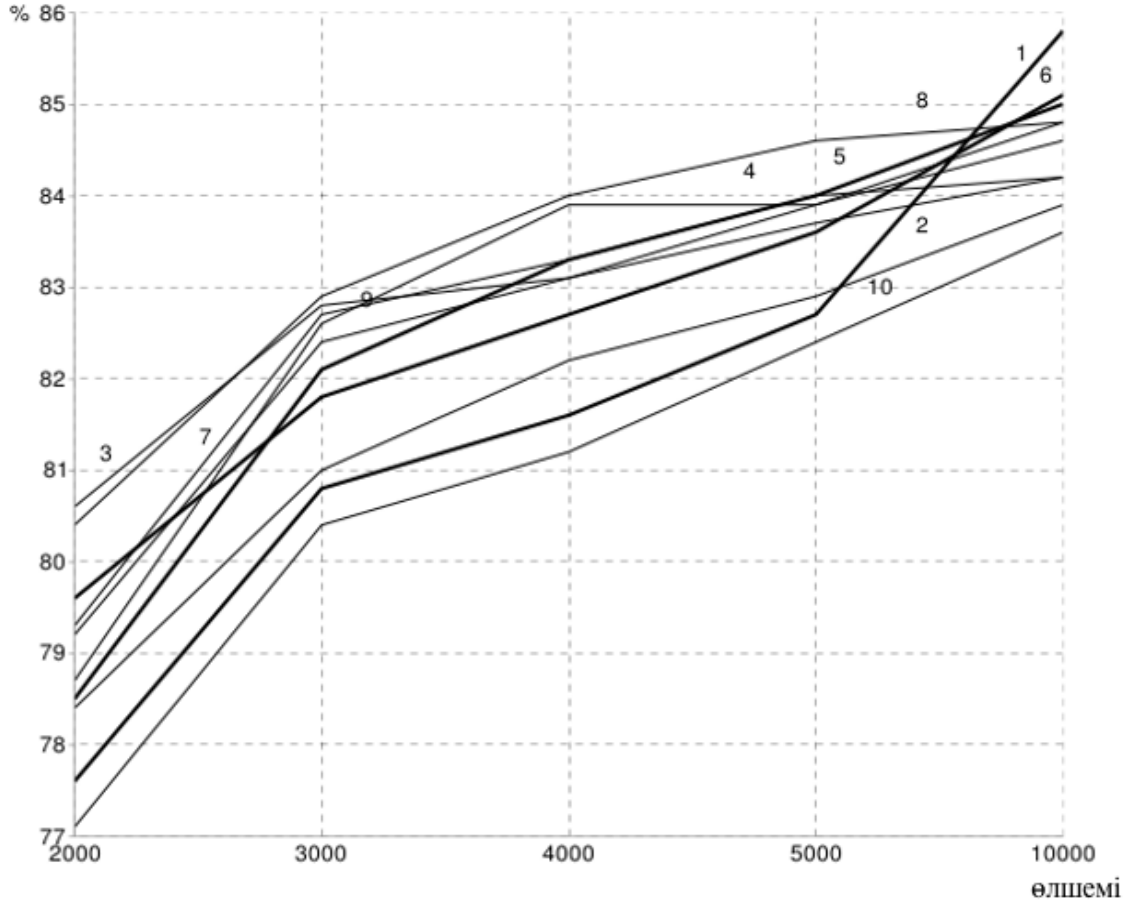
$$1. Weight_r(pw) = \sum_{i=1}^{k} Freq(pw, d_i) \quad (4.8)$$

$$2. Weight_r(pw) = \sum_{i=1}^{k} ln(1 + Freq(pw, d_i) \quad (4.9)$$

$$3. Weight_r(pw) = (\sum_{i=1}^{k} Freq(pw, d_i))|\{d \in r: |Freq(pw, d) > 0\}| \quad (4.10)$$

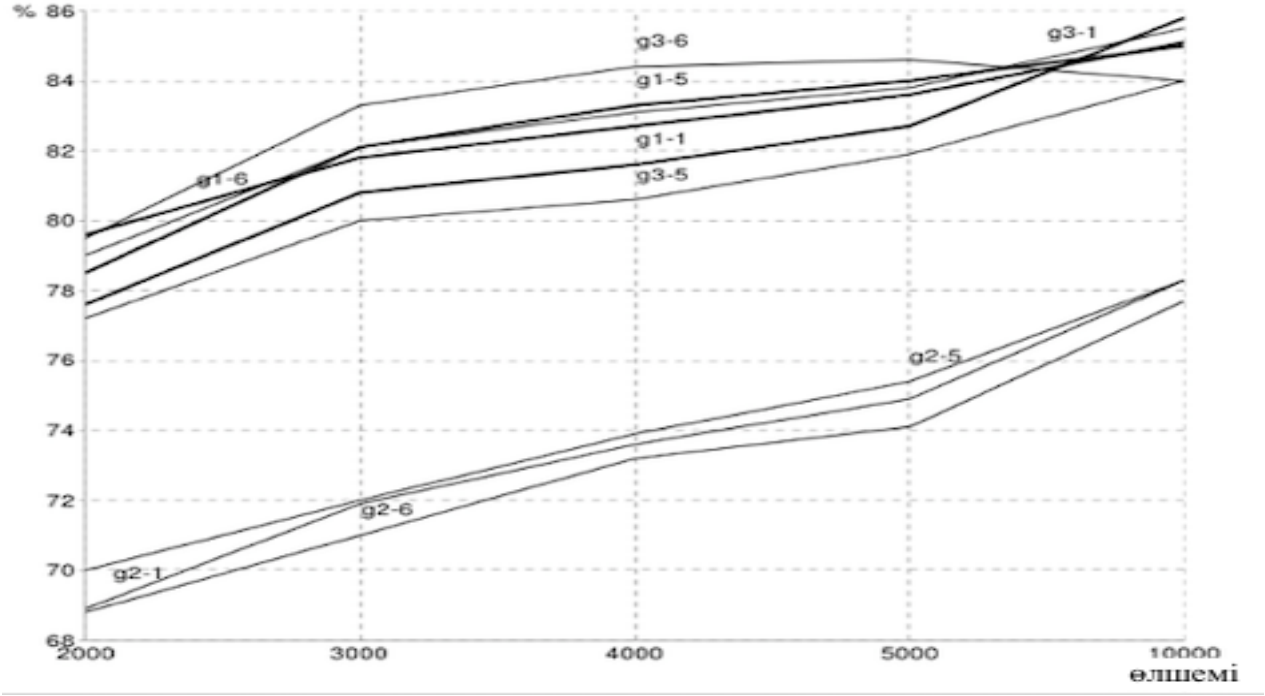
Тестілеу өлшемнің әртүрлі функцияларын қолдану барысында жіктелудің нақтылығын бағалану әрекеті қолданылады. Тестілеудің нәтижесі 2000, 3000, 4000, 5000 және 10000 өлшемдері үшін келтірілген. Бұл терминдерді ерекшелеу алгоритмінің төменгі өлшеміне байланысты нөлдік координатасы бар (15% астам) құжаттардың көптеген саны мен жиіліктің төменгі нақтылығының нәтижесі (70% төмен) тиімсіз болып келеді.

Тестілеудің бірінші сатысында $g1$ функциясы және 8-10 функциясы қолданылған. Нәтижесі сурет 4.3 көрсетілген. Кестедегі функцияның нөмірі оның кестесінің алдыңғы жағында тұрады. Нәтиже (85.8%) 1 номері бойынша функция қолданылған кезде ғана қолжетімді болған. Алайда 10 000 төмен өлшемде басқа функциялармен салыстырғанда оның көрсеткіштері де төмен болып келеді. Айтарлықтай жоғары нәтижені 5 (максимум 85%) және 6 (максимум 85.1%) функциялары көрсетті. Ал қалған функциялар 85% өлшеміне жете алмады. Атап өтейік, 8 және 9 қысқартылған функциялары айтарлықтай жақсы нәтиже көрсетті, ал 8 функциясы 2 000, 3 000, 4 000 өлшемдерінде жақсы нәтиже көрсетті.



Сурет 4.8 - g1 функциясы мен 1-10 функциясын қолдану кезіндегі жіктелудің нақтылығы

Тестілеудің алғашқы сатысынан кейін 2, 3, 4, 7 функциялары, сонымен қатар барлық 8-10 қысқаша функциялар алынып тасталынды. Ары қарай тестілеуге 1, 5, 6, сонымен қатар g_2 және g_3 функциялары қатысты. Тестілеудің екінші деңгейінің нәтижесі сурет 4.4 суретте келтірілген. g_2 функциясының тиімділігі g_1 және g_3 функциясының тиімділігімен салыстырғанда әлдеқайда төмен. g_1 және g_3 функцияларының нәтижесі ұқсас болып келеді, сонымен қатар максималды цифр g_1 және 1 (85.8%), g_3 және 1 (85.5%) функцияларының тіркестеріне жеткен. Ең мықты тіркес ретінде g_1 және 1 тіркесі таңдап алынған, себебі оның максималды ұтымдылығы жоғары, ал тиімсіздігі 2 000-5 000 өлшемдерінен байқалмайды.

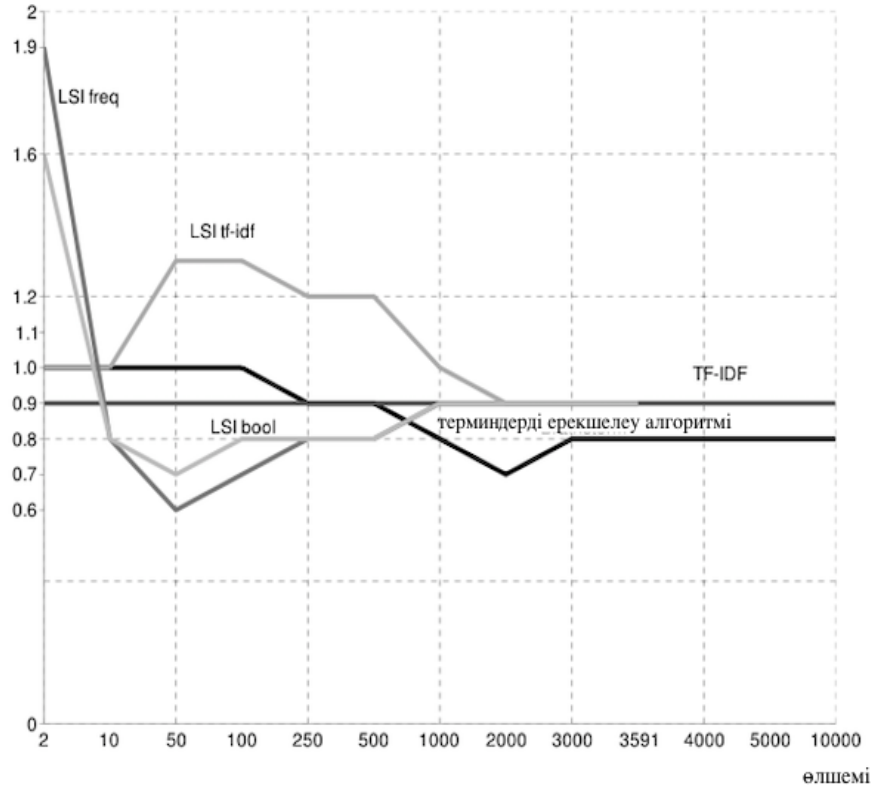


Сурет 4.9 - g1, g2, g3 функциясы мен 1, 5, 6 функциясын қолдану кезіндегі жіктелудің нақтылығы

Осындай жолмен тестілеудің нәтижесі бойынша келесі түріндегі r айдарында pw жұбының өлшемінің функциясын қолдану керектігі шешілген:

$$Weight_r(pw) = \frac{\sqrt{\sum_{d \in r} \ln\left(\frac{\sqrt{Freq(pw,d)}}{\ln\left(\frac{L(d)}{Av(L,D)}+1\right)}+1\right)}}{\ln\left(\frac{|r|}{Av(|\cdot|,R)}+1\right)} \quad (4.11)$$

Кластеризацияның сапасы. Сурет 4.10 кластеризацияның сапасын салыстыру кезіндегі нәтиже көрсетілген. Кластерлерге қатысты азаюдың төменгі мәні алгоритмнің ең жоғарғы тиімділігіне сәйкес келеді. Бұл көрсеткіш бойынша салыстыруға қатысқан алгоритм мөлшермен бірдей нәтижеге жетеді. Сонымен қатар, терминдерді ерекшелеу алгоритмінде LSI алгоритміндегі сияқты кластеризацияның сапасының төмендеуі орын алмайды.



Сурет 4.10 - Терминдерді ерекшелеу алгоритмін қолдану барысындағы кластеризацияның сапасы, LSI және TF-IDF

4.3 кестеде айдардың өлшемі бойынша іріктелген информатика білімінің аумағындағы онтологияны құру барысында алынған алгоритмдердің ішіндегі алғашқы 50 термин келтірілген. Осы онтологияның құру үрдісінің сипаттамасы 4.5 бөлімде берілген.

4.4.3 тұжырым

Бұл бөлімде тақырыптық бөлімдермен берілген мәтіндердің жиынтығынан тұратын терминдерді ерекшелеу алгоритмінің тиімділігін зерттеудің нәтижесі мен бағдарламаны іске асырудың сипаттамасы берілген. Алгоритм 2.3 бөлімде құрастырылған талаптарға сай келеді.

- *Алгоритм ғылыми білімнің аумағындағы шеңбер бойынша берілген әдістер мен тәсілдер, тапсырмалар, бағыттарды сипаттайтын терминдері бар құжаттардың ішінен алынуы керек* кесте 4.1.

- *Ерекшелеп алынған терминдердің нақтылығы мен толық болуы терминдерді ерекшелеу алгоритмінің қазіргі және бұрынғы апробациясының көрсеткіштерінен кем түспеуі керек.*

- *Алготритм үйретуді талап етпеуі керек, немесе оларды өңдеу бойынша үлкен күшті қажет етпейтін ашық дерек көздерінен қажетті үйретушіні таңдап алу мүмкіндігі болуы керек.* Терминдерді ерекшелеу алгоритмі терминдер ерекшеленіп тұратындай үйретуші құжаттардан да, үйретуші терминдерден де тұрмайды. Жалғыз талап терминдерді ерекшелеп алатын құжаттар айдарларға бөлініп тұруы керек.

Кесте 4.1 – Информатика тақырыбындағы терминді ерекшелеу алгоритмімен алынған алғашқы 50 термин

No	термин	No	термин	No	термин
1	data mining	18	experience reports	35	network analysis
2	software engineering	19	data analysis	36	trust management
3	machine learning	20	intrusion detection	37	hoc networks
4	artificial intelligence	21	software systems	38	mining algorithms
5	knowledge discovery	22	wireless communications	39	information integration
6	signal processing	23	knowledge representation	40	query languages
7	data management	24	programming languages	41	data sets
8	computational intelligence	25	data warehousing	42	recommender systems
9	network security	26	requirements engineering	43	reverse engineering
10	wireless networks	27	query processing	44	data quality
11	intelligent systems	28	text mining	45	software architecture
12	software development	29	data integration	46	cognitive radio
13	communication systems	30	digital libraries	47	project management
14	access control	31	software testing	48	data processing
15	formal methods	32	software architectures	49	identity management
16	database systems	33	empirical studies	50	ieee symposium
17	web mining	34	data streams		

- *Алгоритм берілген пәндік аймақтағы орнатулар бойынша сарапшылардың қолымен жасалатын еңбегінің үлкен көлемін талап етпеуі керек.* Алгоритмді пәндік аймаққа орнату үшін сол деңгейлердің әрқайссысында қолданылатын есептік параметрлерді ретке келтіру керек. Белгілі бір білім аумағындағы жұмысқа алгоритмнің үйренісуі үшін бұл параметрлерді өзгертуден басқа әрекеттерді жасаудың қажеті жоқ.

- *Алгоритм үшін дерек көзі қолжетімді болуы керек және жаңартылып отыруы шарт.* Терминдерді ерекшелеу алгоритмі айдарларға бөлінген мәтіндік құжатқа қолданылады.

- *Алгоритм үлгілі архитектурадан құралуы керек.* Терминдерді ерекшелеу алгоритмі төрт үлгілік деңгейден тұрады, олардың әрқайссысында үміткер-сөздердің жұптарына тәуелсіз іріктеу жүргізіледі. Бұл үлгілер тек қана функциялары бойынша ғана емес, сонымен қатар жадыны қолдануы бойынша да тәуелсіз болып табылады.

- *Білімнің белгілі бір аумағындағы алгоритмді автоматты түрде орнату мүмкіндігі болуы керек.* Білімнің белгілі бір аумағындағы алгоритмнің шығыс құжаттарындағы айдарларға бөлінуі мен алдыңғы уақытта атап өтілген параметрлердің автоматты жиынтығынан тұрады. Берілген құжаттарды автоматты түрде немесе автоматтандырылған түрінде айдарларға бөлуге мүмкіндік беретін алгоритмдердің көптеген түрлері бар. Терминдерді ерекшелеу алгоритмінің тиімділігін арттыруға көмектесетін әдістерді атап өтейік.

- *Терминдерді ерекшелеу алгоритмі мен LSI біріктіру.* Терминдерді ерекшелеу алгоритмінің көмегі арқылы терминдерді ерекшелеп алғаннан кейін оларды LSI алгоритміне енгізуге жіберуге болады (яғни барлық сөздердің ішіндегі матрицаның орнына терминдердің ішіндегі матрицаны қарастыру). Болжам бойынша бұл жіктеудің нақтылығын арттыруға барынша көмектеседі.

- *Терминдерді ерекшелеу алгоритмінің үштік, төрттік және тағы сол сияқты сөздерге кеңейтілуі.* Оны екі түрлі тәсіл арқылы іске асыруға болады: не алгоритмнің өзін қайта өңдеу, немесе таңдап алынған терминдерге сөздердің жұбынан құралған барынша ұзын тізбектерді қолдану арқылы сондай күйінде қолдану (мысалы, сәйкес келетін сөздер бойынша бірнеше жұптарды біріктіру).

- *Лингвистикалық тәсілдерді қолдану***.** Жоғарыда атап өткендей, терминдерді ерекшелеу алгоритмі алғашқыда қолданылатын лемматизациядан басқа (алгоритмді қолданғанға дейін орындалуы керек) ешқандай лингвистикалық әдістерді қолданбайды. Лингвистикалық әдістер алгоритмнің нәтижесін жақсартуға қабілетті деген болжам жасауға негіз бар. Ең қарапайым кеңейтілу жолы құжаттардың ішінен стоп-сөзінің алғашқы сүзбесін (фильтрация) іске қосу болып табылады. Басқа да қызықты бағыттар – жұптарды іріктеу деңгейінде синтаксистік талдау ағашын қолдану. Ағаштың бір бұтағында орналасқан жұптарды таңдау әрекеті алгоритмнің жұмыс уақытын азайтып, оның тиімділігін арттыруы мүмкін.

- *IDF жоғарғы мәні бар дара сөздер базисіне енгізулер жасау.* Бұл осы сөздермен іріктелініп алынған және базистің өлшемін қысқартатын біраз көлемді жұптардың базисінен алып тастауға мүмкіндік береді.

- *Толықтай кеңістік өлшеміне тәуелді алгоритмдердің параметрлерін анықтаушы тұтынушының функциясын іске қосу мүмкіндігі негізгі мақсат болып табылатын алгоритмнің модификациясы.* Алгоритм тиімділігінің қорытындысының нәтижесіндегі көрсеткіш бойынша көп жағдайларда жіктелудің нақты болуының жоғарғы көрсеткіші әртүрлі өлшемдерде параметрлердің мәні әртүрлі болған жағдайда орын алады.

- *Айдардың қуаттылығына байланысты әртүрлі жұмыс барысындағы алгоритмнің тиімділігінің артуы.* Терминдерді ерекшелеу алгоритмінің тиімді жұмыс істеуі үшін шығыс айдарлары мөлшермен құжаттардың санына тең болуы керек. Тестілеудің көрсеткіші бойынша айырмашылықтар 2-3 есе болады, бірақ айырмашылық болған жағдайда ерекшеленген терминдердің сапасы айтарлықтай төмендейді. Айдардың қуаттылығы бойынша әртүрлі айдарлармен алгоритм жұмысының тиімділігін арттыру үшін оның модификациясы қажет.

Терминдерді ерекшелеу алгоритмі өзіне қойылатын талаптарға сай келеді және мәтіндік құжаттардан алынуы керек терминдердегі, басқа да жүйелердегі сияқты, ғылыми ақпараттың қорытындысы мен есептеу жүйесінің сәйкес келетін үлгісіндегідей қолданылуы мүмкін.

4.5 Онтологияны құрастыру алгоритмінің тиімділігін зерттеу, бағдарламалық іске асыру

Берілген бөлімде ғылыми конференциялардың негізінде ғылыми білім аумағындағы онтологияны құрастыру алгоритмінің тиімділігін зерттеудің

нәтижесі көрсетілген. Конференция анонстарының шығыс жиынтығы ретінде 13098 құжаттан тұратын порталдың базасы WikiCFP қолданылған. Құжаттарды айдарларға бөлу [50-55] олардың сайттар арқылы үлесіне енгізген белгілердің негізінде орын алған. Tags белгілерінің жиыны 6643 бөлшектен тұрады. Құжаттарды айдарларға бөлу үрдісі іске асыратын Mathematica 5 бағдарламалық жиынтығының FindClusters функциясының көмегі арқылы қарастырылып жатқан құжаттардың көпшілігі 7 айдарға бөлінген, олар мынандай мөлшердегі құжаттардан тұрады: 9631, 917, 519, 200, 496, 794 және 283. Білім аумағы ретінде тесттік сынақтарда информатика аймағы қолданылды кесте 4.2.

Кесте 4.2 - Фильтрациядан өткізудің бірінші деңгейінде алынып тасталған алғашқы 20 термин

No	термин	No	термин
1	doctoral symposium	11	tool support
2	multimedia databases	12	vision papers
3	experience papers	13	ieee 2-column
4	software processes	14	well-marked appendices
5	novel data	15	graph mining
6	process models	16	security requirements
7	preserving data	17	copies due
8	original technical	18	information flow
9	intelligent user	19	camera-ready
10	distributed software	20	substantially overlap

Фильтрациядан кейін 1103 термин таңдап алынған. Нақты мөлшері 73.5% құрады, толықтай алғанда – 81%. Атап өтейік, толықтай өлшем ретінде [56-60] ұсынылған жергілікті толық өлшемнің қолданысы пайдаланылады.

Фильтрациядан өткізудің екінші өлшемін қолданғаннан кейін 1103 терминнен 843 термин таңдап алынған. Нақты мөлшері 77% құрады, толықтай алғанда – 63.6% кесте 4.3.

Кесте 4.3 – Фильтрациядан өткізудің екінші деңгейінде алынып тасталған алғашқы 20 термин

No	термин	No	термин
1	access control	11	database technology
2	experience reports	12	model-driven engineering
3	recommender systems	13	model-driven development
4	web search	14	symposium series
5	collaborative filtering	15	software process
6	empirical software	16	web-based services
7	must conform	17	autonomous agents
8	personal information	18	pattern mining
9	modeling languages	19	multimedia communication
10	digital rights	20	services security

Кесте 4.4 терминдерді фильтрациядан өткізу мен ерекшелеудің тиімділігінің қорытындысы келтірілген. Нақтылық пен толықтығын анықтау барысында [61-66] сарапшының көзқарасы бойынша информатика аумағындағы қолданысқа тән сөздердің жұптары термин деп есептелген.

Кесте 4.4 – Терминдерді фильтрациядан өткізу мен ерекшелеу деңгейінің тиімділігін бағалау

No	алгоритм қадамы	ерекшеленген термин	нақтылық	жергілікті толықтылық	F-өлшем
1	терминді ерекшелеу	1793	70.5%	-	-
2	фильтрация-1	1103	73.5%	81%	77.1%
3	фильтрация-2	843	77%	63.6%	69.6%
2-3	жалпы фильтрация	843	77%	51.7%	61.9%

Біріккен қатынасты анықтау барысында 381501 ішінен 3771 жұп ерекшеленіп алынған. Орташа есеппен алғанда әрбір термин басқа төрт терминмен байланысты болып келеді. Одан бөлек, 85 иерархиялық қатынас (нақтылығы 89.4% құрады) пен 135 деңгейлік қатынас (нақтылығы 85.2% құрады) [67-73].

Терминдерді орыс тіліне аудару сатысында 874 терминнен (45.9%) 401 термин үшін Википедияда мақала табылған. Олардың ішіндегі 212 термин үшін (24.2%) онтологияға енгізілген орысша эквиваленттер табылған.

Терминдерді қазақ тіліне аудару сатысында 704 терминнен (40.1%) 321 термин үшін Википедияда мақала табылған. Олардың ішіндегі 169 термин үшін (19.2%) онтологияға енгізілген қазақша эквиваленттер табылған.

Алгоритм жұмысының нәтижесінде 61 класс, 1086 дана және 4203 қатынастан тұратын информатиканың онтологиясы құралған. Оның фрагменті сурет 4.8 келтірілген.

4.5.2 бөлім бойынша тұжырым

Берілген бөлімде Ғаламтордағы ақпараттар мен ғылыми конференциялардың негізіндегі ғылыми білім аумағындағы онтологияның құрылу алгоритмінің тиімділігін зерттеудің нәтижесі мен бағдарламаны іске асырудың сипаттамасы келтірілген. Алгоритмнің 2.4 бөлімде құрастырылған талаптарға сай келетіндігін көрсетілді.

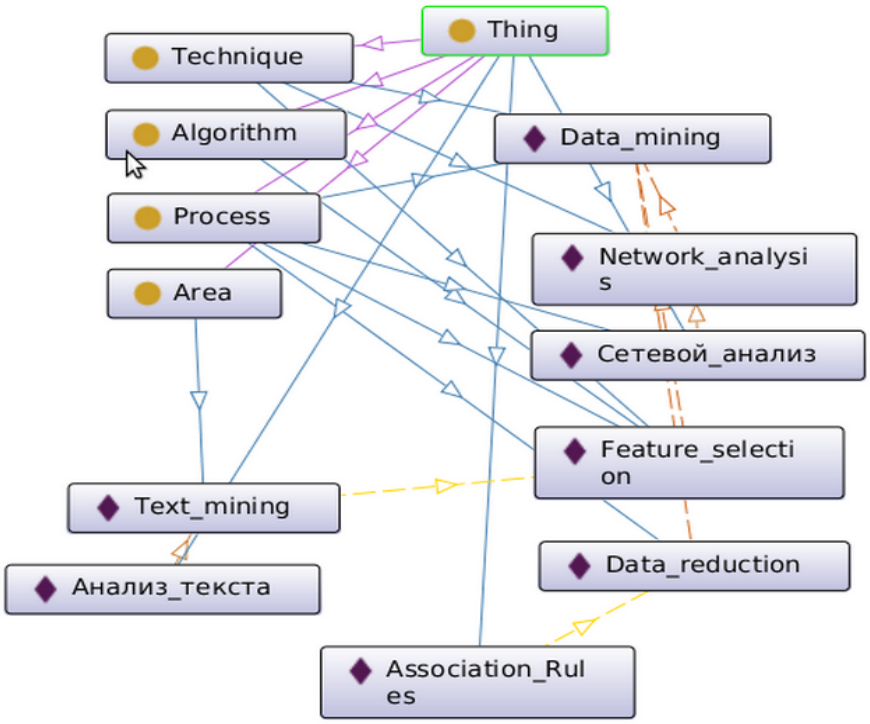
Алгоритм терминдердің арасындағы иерархиялық және ассоциативті қатынасты көрсетуі керек. Онтологияны құрайтын алгоритмнің құрамына терминдердің арасындағы иерархиялық және ассоциативті қатынасты ерекшелейтін үлгі енген. Алгоритмнің көмегі арқылы информатиканың пәндік аумағы үшін 3771 ассоциатитівті қатынас және иерархиялық деп қарастыруға болатын 135 категориялық және 85 иерархиялық термин берілген.

Кесте 4.5 - Терминдердің арасындағы қатынасты анықтаушы деңгейлерге баға беру

Байланыс типі	Ерекшеленген тип	Нақтылық
Ассоциативті	3771	-
Иерархиялық	85	89.4%
Санатты	135	85.2%

Ғылыми білім аумағындағы онтологияның даналарының жиынтығын құрайтын терминдерді шығарушы ретінде ғылыми конференциялардың анонсы қолданылады. Конференциялар кейде жылдап, кейде бірнеше айларға кешігіп жататын мақалалар, журналдардан тиімді түрде ерекшелей түсетін ғылымның бөлімінің қазіргі жағдайын көрсетеді. Конференциялар туралы ақпарат білім аумағының барлық деңгейі бойынша дамуын, тура мағынада айтқанда, бақылап отыруға мүмкіндік туғызу үшін үнемі толықтырылып отырады. Конференциялардың тақырыбының тізімі мен қабылданған жұмыстардың тізімдері олардың нақтылығын бақылап үшін сарапшылардың өздері толтырады.

- *Алынатын терминдер мен қарым-қатынастың толықтығы мен нақтылығы онтологияны құрайтын алгоритмдердің қазіргі және бұрынғы сапасынан кем түспеуі керек.* Қорытындының нәтижесіндегі көрсеткіш бойынша онтологияны құрайтын алгоритм жоғарғы деңгейдегі нақтылығы және толық болуының арасындағы байланыс пен терминдерді ерекшелеуге мүмкіндік береді.



Сурет 4.11 - Информатика аймағындағы онтологияны құру алгоритмнің көмегімен құралған фрагмент

Ғылыми конференциялардың жарнамасы ғылыми білім аймағы жайлы іріктелген және толық ақпараттан тұрады. Егер ғылыми бағыт дамитын болса,

білім аймағының айтарлықтай толыққандылығын дәлелдейтін, оған арналған конференция бар деген сөз. Ғылымның ірі бөлімдері бойынша, мысалы деректер базасы, жылына бірнеше конференциялар өткізіледі. Конференциялар туралы ақпараттан білім аумағындағы кілттік түсінік пен олардың арасындағы байланысты ерекшелеуге болады, яғни білім аймағындағы онтологияны толықтыру керек. Конференциялар ғылыми бөлімдердің қазіргі жағдайын бейнелейді, яғни, олар туралы ақпараттың қорытындысы әлемдік масштабтағы ұйымдар мен әртүрлі ғалымдардың үлестері, ғылыми бағытты бағалауға мүмкіндік береді.

Ақпарат таратушы ретінде CFP құжаттарын қолданудың тағы бір артықшылығы халықаралық конференциялар туралы ақпараттардан тұратын сілтемелердің тізіміндегі хат әдетте қарапайым құрылымнан тұратындығы болып табылады. Конференциялардың веб-сайттары үшін де аз мөлшерде ғана дұрыс болып келеді [74-76].

- *Алгоритм үйретуде көп күш жұмсамай ашық дерек көздерінен алынған қажетті ақпаратты ала алатын мүмкіндігі бар болуы керек.* Іздеу жүйелері тауып беретін Ғаламтордағы ізденістің нәтижесінен алынған ақпараттарды қолданудың есебінен онтологияны құраушы алгоритм үйретуді қажет етпейді. Енетін деректерге қойылатын алгоритмнің жалғыз талабы терминдерді ерекшелеп алатын құжаттар айдарларға бөлініп тұруы керек.

- *Алгоритм берілген пәндік аймақтағы орнатулар бойынша сарапшылардың қолмен жасалатын еңбегінің үлкен көлемін талап етпеуі керек.* Алгоритмді пәндік аймаққа орнату үшін сол деңгейлердің әрқайссысында қолданылатын есептік параметрлерді ретке келтіру керек. Белгілі бір білім аймағындағы жұмысқа алгоритмнің үйренісуі үшін бұл параметрлерді өзгертуден басқа әрекеттерді жасаудың қажеті жоқ.

- *Алгоритм үшін дерек көзі қолжетімді болуы керек және жаңартылып отыруы керек.* Ғылыми конференциялардың хабарландыруларының сілтеме тізімдері мен WikiCFP сайтындағы ақпарат қолжетімді және үнемі жаңартылып отырады. Бұл жағдай ғылыми білім аймағының үлгісін белсенді қалыпта сақтап тұру үшін талап етілетін кезеңдер бойынша онтологияны құру алгоритмін қолдануға мүмкіндік береді.

- *Алгоритм модульды архитектурадан құралуы керек.* Онтологияны құру алгоритмі жеті үлгілік деңгейден тұрады. Онтологияның терминологиялық бөлігін құрайтын бірінші деңгей диссертацияда толығымен сипатталған. Бұл деңгейде құрылған онтология түсінігі ғылыми білімнің аймағындағы онтологияны құру үшін қолданылуы мүмкін. Екінші деңгейде 4.4.3 бөлімде көрсетілгендей үлгілік архитектураға ие терминдерді ерекшелеу алгоритмі қолданылады. 3-7 деңгейлерде терминдердің фильтрациялануы мен олардың арасында әртүрлі қатынас орнатылады. Бұл деңгейлер бір-біріне тәуелсіз болады.

- *Білімнің белгілі бір аймағындағы алгоритмді автоматты түрде орнату мүмкіндігі болуы керек.* Білімнің белгілі бір аймағындағы алгоритмді шығыс құжаттарының айдарларға бөлінуі мен алдыңғы уақытта атап өтілген параметрлердің автоматты жиынтығынан тұрады. 2.3 бөлімде көрсетілгендей

ғылыми конференциялардың хабарландыруының айдары шынайы айдарларды туғызатын сілтемелердің бірнеше тізімдерін қолдану мен WikiCFP порталын қолданушылардың белгілеуін пайдалану арқылы оңай орын алуы мүмкін.

Білім аймағындағы онтологияны құру алгоритмінің бірнеше кемшіліктерін атап өтейік.

- *Терминдердің түрлеріне қойылатын шектеулер*. Терминдерді ерекшелеуде үш немесе одан да көп сөзден тұратын терминдерді ерекшелеу үшін өзгертілген.

- *Терминдердің арасындағы байланысты анықтау барысында кеңейту ескерілмейді.* Қорытынды екі терминнің арасындағы байланысты анықтауда сөйлемнің жартысы ғана ескеріледі. Маңызды ақпаратты бірінші терминнің сол жағында немесе екінші терминнің оң жағында орналасқан сөйлемдердің жартысының мазмұнында сақтауға болатын еді.

- *Атаулы байланыс түрлерін ерекшелеудің айтарлықтай толық болмауы.* Байланысын анықтаушы алгоритмді қолданудың нәтижесінде 874 терминдердің арасынан 85 иерархилық және 135 категориялық байланыс анықталған. Осындай жолмен әрбір термин үшін орташа есеппен алғанда 0.25 қатынастан келген, ол айтарлықтай жеткіліксіз деуге болады.

- *Атауы бар қатынастарды ерекшелейтін типтің аз мөлшері.* Бұл түрінде алгоритм иерархиялық қатынас пен деңгейлік қатынасты ерекшелеуге мүмкіндік береді. Бірігу қатынасының басқа түрлері ерекшеленбейді.

- *Есептеудің жоғарғы қиындығы***.** Онтологияны құрайтын алгоритмнің тесттік сынағының нәтижелері 3 теорема арқылы дәлелденген оның есептеу қиындығын аналитикалық тұрғыдан бағалаумен байланысты болып келеді. Соған байланысты әрбір термин үшін іздеу жүйесіне деген сұраныстың үлкен көлемін орындау қажет, онтологияны құру алгоритмі есептеудің жоғарғы қиындығынан тұрады. Сонымен қатар жергілікті есептеу торы арқылы орын алатын операциялар тәртіп бойынша тезірек жұмыс істейді. Ақпараттық жүйелерде тәжірибеде ұсынылған алгоритмді қолдануға мүмкіндік беретіндей, онтология салыстырмалы түрде сирек құралады және қайтадан түзеледі.

Онтологияны құрайтын заманауи алгоритмдермен салыстырғанда онтологияны құру алгоритмі мынандай артықшылықтарға ие:

- шығыс деректеріне қойылатын оңай талаптар;
- білім аймағындағы терминдерді автоматты түрде ерекшелеу;
- атауы бар қатынастарды ерекшелеудің жоғары дәлдігі;
- алгоритм модификациясыз ғылыми білімнің кез-келген аймағына және білімнің басқа да көптеген аймақтарына қолданылуы мүмкін екендігін ескере отырып, оның әмбебап екендігін айтуға болады;
- сарапшылардың қолмен жасайтын еңбегінің қажеттілігінің орын алмауы. Ғаламтор мен ғылыми конференциялардың жарнамаларының негізіндегі ғылыми білім аймағындағы онтологияның құру алгоритмінің тиімділігін зерттеудің нәтижесінің қорытындысы алгоритм тәжірибеде жақсы фактілерді берді. Онтологияны құру алгоритмі қойылған талаптарға сай келді. Берілген диссертация тапсырмалары мен қатар ғылыми білімнің сәйкес қандай да бір басқа аймағында қолдануға болады.

4.6 Бағдарламалық интерфейсті қолданудың тиімділігін бағалау

Біркелкі фрагментті анықтау бойынша бағдарламамен қамтамасыз етудің ерекшелігі тұтынушылық физиология мен психология ескерілетін HCD (ағылшын тілінен Human Centered Design, адамға бағытталған жоба) әдістеменің кейбір негізгі бөлімдері сақталынған жұмысты жобалау барысындағы қолданыстық интерфейс болып табылады.

Бағдарламалық интерфейсті қолданудың қолайлылығын анықтау үшін негізіне сапаның қасиеті енгізілген бағалаудың әртүрлі әдістері қолданылуы мүмкін. Сонымен қатар қолданыстың қолайлылығы функционалдық талаптың үлгісі бола алмайды, соған сәйкес ол турасынан өлшенбеуі мүмкін, бірақ мысалы, жүйенің экплуатациялық қолайлылығына қатысты мәселелер туралы хабарламалардың саны сияқты атрибуттар мен түзу емес өлшемдердің тәсілі бойынша анықталуы керек.

Сол арқылы кез-келген жобаның шеңберінде басқаруға болатын қолданыстың қолайлылығын анықтау үшін сапаның тәсілдері мен сипаттамалардың белгілі бір жиынтығы жоқ. Сапаның қасиеттері үйренушілер шешуі керек тапсырмалардың жиынтығы мен қолданыс контекстін ескерместен қалыптасады, қашықтықтан оқытудың бағдарламалық тәсілдеріне қойылатын талаптардан тыс анықталады. Берілген зерттеудің шеңберінде бағдарламалық қамтамасыздандырудың интерфейсін қолданудың қолайлылығын бағалау үшін тұтынушы интерфейсінің мәдениеті мен тұтынушының қателіктерінен қорғану, қолданыстың қарапайымдылығы, танымалдылық сияқты 4 сипаттама қолданылған.

Танымалдылық – тестілеуге қатысушылар терезе қосымшалармен жұмыс істеудің тек бастапқы дағдысын ғана біледі деп күтіледі.

Қолданудың қарапайымдылығы – тестілеуге қатысушылар гетерогенді деректерді басқарумен жұмыс істеуде тәжірибесі болмағандықтан, мысалы, қателік жіберу бұл сипаттамаға барынша ықпал етуі мүмкін.

Тұтынушының қателіктерінен қорғану – өзара әрекеттесу жүйесі тестілеуге қатысушыларды қате жіберуден сақтандыруы керек.

Тұтынушы интерфейсінің мәдениеті – интерфейсті тестілеуге қатысушыларды қызығушылығын және жүйені қолдану барысында қанағаттанарлықтай жағдай туғызуы керек. Жоғарыда сипатталған 4 сипаттама нәтиже беруші кестеде сипатталатын сапаның белгілі бір қасиетіне ие.

Тестілеуге қатысушылар ретінде Халықаралық ақпараттық технология университетінің «Компьютерлік инженерлік» кафедрасының 10 оқытушысы мен «Бағдарламалық қамтамасыздандыру және есептеу техникасы» мамандығы бойынша 20 студент алынған. Барлық бағаны жинақтағаннан кейін 30 қатысушының ішінен барлық нәтиже орташа мәнге дейін жеткен.

Бағдарламалық интерфейсті қолданудың қолайлылығын бағалаудың берілген нәтижесі бір санның көмегімен толықтай суретті көрсетпейді және мүмкін болатын мәселені анықтай алмайды.

Кесте 4.6 – Бағдарламалық интерфейсті бағалау

Сипаты	Сапасының қасиеттері	Нәтиже
Танымалдық	Бір дегеннен сәтті орындалған функция саны	84%
	Бірінші рет қолдануға жіберілген уақыт	28 сек.
	Қолданыстағы функция түсініктілігіне қолданушының бағасы	92%
Қолданудың қарапайымдылығы	Кеңессіз өткен нәтижелі уақыт	14 мин.
	Жүйе жасаушыларына қатысты сұрақтар саны	4
	Қолданушыр тарапынан қолдану қарапайымдылық бағасы	87%
Тұтынушының қателіктерінен қорғау	Қате туралы хабарламалардың түсініктілігі	74%
	Қате туралы хабарламалардың ақпараттылығын пайдаланушының бағасы	80%
Тұтынушы интерфейсінің мәдениеті	Пайдаланушы интерфейсі эстетикасына қанағаттанушылық бағасы	92%

4.7 Бұл аймақтағы ары қарай жүргізілетін зерттеу

Жүйенің ары қарай дамуы университеттің бөлімдерінің ғылыми жұмыстарының салыстырмалы тақырыптық қорытындылау жүйесінің жасалуы, олардың енгізген деректерінің негізінде автоматтандырылған тәртіптегі ұйымның серіктестерінің ғылыми балдарының есебі мен жылдық ғылыми есепті құраушы қызметті іске асырудан тұрады. Ол үшін берілген диссертацияда сипатталған үлгілердің көмегімен жүйені кеңейту жоспарлануда, жекелей алғанда, жүйенің сақтау қоймасындағы деректер, мақалалардың тақырыптық жіктелу үлгісі, Ғаламтордағы ақпаратты іздеу үлгісі, онтологияның көмегімен білім аумағындағы семантикалық үлгіні құраушы үлгілерді жатқызуға болады.

4 бөлім бойынша тұжырым

Бұл бөлімде ғылыми білім аумағындағы онтологияның құрылуы мен берілген тақырыптық бөлімдермен бірге мәтіндер жиынтығындағы терминдерді ерекшелеу алгоритмінің тиімділігін зерттеудің нәтижесі көрсетілген. Жасалынған қорытынды нәтижесінің көрсеткіші бойынша алгоритм оған қойылатын талаптарға сай және ғылыми ақпараттың қорытындысы мен есептеу жүйесінің үлгісіне сәйкес қолданылуы мүмкін. Алгоритмдер тәжірибелік білімнің аймағындағы жеке мәселеге қатысты тәуелсіз болып келеді. Олар пәндік аймақтың онтологиясын құрып, құжаттардың ішінен терминдерді ерекшелеп алу қажеттілігі туындайтын басқа жүйелерде де қолданылуы мүмкін.

ҚОРЫТЫНДЫ

Бұл диссертацияда ғылыми ақпараттарды басқару жүйесінің құрылу тәсілі мен әдістерінің сипаттамалары берілген. Әрекет етудің теоретикалық негізі онтология болып табылады. Жүйенің авторы ұсынған нұсқа сұраныстарды орындау үлгісі, онтологияны құрау мен ғалымдардың ғылыми жұмыстарының нәтижесі туралы деректермен жүйеге енгізілген байланысты орнатушы үлгі, ақпараттарды жүктеуші үлгі, білім аумағындағы формальды үлгіні құраушы үлгі сияқты үлгілерді өз құрамында сақтайды.

Зерттеу жұмысында ұсынылған онтологияға негізделген ақпараттық жүйелерді құрауға ұсынылған әрекет зерттеудің ғылыми жұмысының қорытындысын есепке алу үшін ғана емес, сонымен қатар басқа да маңызы бар тәжірибелік тапсырмаларды шешу үшін қолданылады.

Зерттеу барысында келесідей нәтижелер алынды.

- Ғылыми ұйымның қызметінің нәтижесін сипаттауда пәндік аймақ зерттеуінің нәтижесі негізінде онтологияны қолдана отырып гетерогенді құрылымы бар үлкен деректермен жұмыс жасауда қолданатын алгоритм әзірлемесі жасалды

- Онтология мен SPARQL тілін қолдану арқылы оның бүкіл өмір сүру циклындағы жүйенің кодының тиімділігі үшін қосымша мүмкіндіктер мен есептеудің кепілдігін туғызатын жүйеге деген сұраныстардың формальды сипаттамасы ұсынылған.

- Ғаламтордағы іздеу жүйесіндегі ақпараттарды қолдануға, ғылыми конференциялардың жарнамаларындағы терминдерді ерекшелеуге негізделген ғылыми білім аумағындағы онтологияны құру алгоритм жасалынған. Оның бағдарламасының іске асырылуының күрделілігін сипаттайтын аналитикалық баға алынған.

- Тақырыптық бөлімдермен берілген мәтіндердің жиынтығынан термин-жұп сөздерді ерекшелеу алгоритмі жасалынған. Автор ұсынған алгоритмнің құрамында аударудағы терминнің салмағының негізгі қызметі оған қойылатын талаптарды қанағаттандырады. Алгоритмнің бағдарламалық іске асырылуының күрделілігін анықтауды сипаттайтын аналитикалық баға анықталған.

ПАЙДАЛАНЫЛҒАН ӘДЕБИЕТТЕР ТІЗІМІ

1 Қазақстан Республикасы Үкіметінің №827 Қаулысымен «Цифрлық Қазақстан» 2017 жылғы 12 желтоқсанда бекітілген мемлекеттік бағдарлама.

2 Turner V., Gantz J.F., Reinsel D., et al. The Digital Universe of Opportunities: Rich Data and the creasing Value of the Internet of Things. IDC white paper. - 2014. // http://idcdocserv.com/1678 (қаралым күні: 31.08.2021).

3 Global Enterprise Big Data Trends: 2013. Microsoft corp. survey. – 2013 // https://news.microsoft.com/download/presskits/bigdata/docs/big- data_021113.pdf (қаралым күні: 31.08.2021).

4 Иванов П.Д., Вампилова В.Ж. Технологии BigData и их применение на современном промышленном предприятии // Инженерный журнал: наука и инновации. - М.: МГТУ, 2014. - №8(32). – С. 10.

5 Семенов Ю.А., Большие обьемы данных (BigData) // http://book.itep.ru/4/7/big_data.html (1.1.2) (қаралым күні: 31.08.2021).

6 Zikopoulos P., Parasuraman K., Deutsch T., Giles J., Corrigan D. Harness the power of big data The IBM big data platform. - New York: McGraw Hill Professional, 2012 //http:// books.google.com/books?id=HhSONOxOCQOC (дата обращения: 31.08.2021).

7 Biehn Neil. The Missing V's in Big Data: Viability and Value // https://www.wired.com/insights/2013/05/the-missing-vs-in-big-data-viability-and-value/ (қаралым күні: 31.08.2021).

8 Big Data Analytics Landscape. -2019 // https://www.learnbigdatatools.com/big-data-analytics-landscape-2019 (қаралым күні: 31.08.2021).

9 Big Data Executive Survey // Executive Summary of Findings. - 2017 // http://newvantage.com/wp-content/uploads/2017/01/Big-Data-Executive-Survey-2017-Executive-Summary .pdf (қаралым күні: 31.08.2021).

10 MapReduce: Simplied Data Processing on Large Clusters Jeffrey Dean and Sanjay Ghemaw. To appearing. - OSDI, 2014. - P. 13.

11 NoSQL Databases. – 2016, september // http://www.nosql-database.org (қаралым күні: 31.08.2021).

12 Афонин С.А. и др. Интеллектуальная система тематического исследования научно-технической информации (ИСТИНА) / под ред. Академика В.А. Садовничего. – М.: Издательство МГУ, 2014. – 262 с.

13 Данилова Т.С., Зелепухина В.А., Бурмистров А.С., Тарасевич Ю.Ю. Информационно-аналитическая система для сбора, хранения и анализа научной и наукометрической информации. Руководство пользователя / под ред. Ю.Ю. Тарасевича. – Астрахань: ООО «Типография Новая Линия», 2014. – 191 с.

14 About Pure // Elsevier // https://www.elsevier.com/solutions/pure (қаралым күні: 31.08.2021).

15 Bibster – A Semantics-Based Bibliographic Peer-to-Peer System/Peter Haase, Jeen Broekstra, Marc Ehrig et al. The Semantic Web – ISWC 2004 / ed. by Sheila A McIlraith, Dimitris Plexousakis, Frank van Harmelen // Lecture Notes in Computer Science. – Berlin: Springer; Heidelberg, 2004. - Vol. 3298. - P. 122–136.

16 The swrc ontology - semantic web for research communities. York Sure, Stephan Bloehdorn, Peter Haase et al // Proceedings of the 12th Portuguese Conference on Artificial Intelligence - Progress in Artificial Intelligence (EPIA 2005) LNCS. - Covilha: Springer, 2005. - Vol. 3803. - P. 218–231.

17 Lassila Ora, Swick Ralph R. Resource Description Framework (RDF) Model and Syntax Specification // World Wide Web Internet And Web Information Systems. - 1999. - Vol. -P. 1–54.

18 Kruk, Sebastian Ryszard. JeromeDL - Adding Semantic Web Technologies to Digital Libraries. Sebastian Ryszard Kruk, Stefan Decker, Lech Zieborak. Database and Expert Systems Applications / ed. by Kim Viborg Andersen, John Debenham, Roland Wagner // Lecture Notes in Computer Science. – Berlin: Springer; Heidelberg, 2005. - Vol. 3588. - P. 716–725.

19 Brickley Dan. FOAF Vocabulary Specification. – 2010 // http://xmlns.com/foaf/spec/ (дата обращения: 31.08.2021).

20 Ponomareva Natalia., Gomez Jose, Pekar Viktor. AIR: A Semi-Automatic System for Archiving Institutional Repositories. Natural Language Processing and Information Systems / ed. by Helmut Horacek, Elisabeth M´etais, Rafael Mun˜oz, Magdalena Wolska // Lecture Notes in Computer Science. – Berlin: Springer; Heidelberg, 2010. - Vol. 5723. - P. 169–181.

21 Коськин А.В., Ужаринский А.Ю., Титенко Е.А. Место web-сервисов в системе управления образовательным учреждением // Известия Юго-Западного государственного университета. – 2013. – №1 (46). – С. 70-75.

22 ГОСТ Р ИСО 9241-210–2012. Эргономика взаимодействия человек-система. Человеко-ориентированное проектирование интерактивных систем. – Введ. 2018-01-12. – М.: Стандарт информ, 2018. – Ч. 210. - 36 с.

23 Gruber Thomas R. A translation approach to portable ontology specifications // Knowledge Acquisition. - 1993. - Vol. 5, №2. - P. 199–220.

24 McGuinness D., Harmelen F. Van OWL Web Ontology Language Overview // Tech. Rep. W3C. - 2004. - №99. – P. 289-305

25 Horrocks Ian. DAML+OIL: a Description Logic for the Semantic Web // IEEE Data Engineering Bulletin. - 2002. - Vol. 25, №1. - P. 4–9.

26 Fareh Messaouda, Boussaid Omar, Challal Rachid. Multi-level Metadata Integration System: XML, RDF and RuleML // World Academy of Science. Engineering and Technology. – 2012. - №62. – P. 845-850.

27 Астраханцев Н.А. Методы и программные средства извлечения терминов из коллекции текстовых документов предметной области: дис. ... канд. физ.-мат. наук: 05.13.11. – М., 2014. – 148 с.

28 Бова В.В., Заруба Д.В., Курейчик В.В. Эволюционный подход к решению задачи интеграции онтологий // Известия ЮФУ. Технические науки. - 2015. - №6(167). - С. 41-56.

29 Глушков Н.А. Анализ методов тематического моделирования текстов на естественном языке // Молодой ученый. – 2018. – №19. – С. 101-103.

30 Долгий А.И., Колоденкова А.Е., Ковалев С.М. Проблемы и методы слияния разнородных данных в гибридных интеллектуальных системах // Гибридные и синергетические интеллектуальные системы: материалы IV

Всероссийской Поспеловской конференции с международным участием. – Калининград: БФУ им. И. Канта, 2018. – С.181-187.

31 Голенков В.В. Онтологическое проектирование гибридных семантически совместимых интеллектуальных систем на основе смыслового представления знаний // Онтология проектирования. – 2019. – Т. 9, №1(31). – С. 132-151.

32 Волкова И.А., Головин И.Г. Лингвистический процессор русского языка: анализ устойчивых словосочетаний // Научные труды SWorld. – 2015. – Т. 2, №4 (41). – С. 36-46.

33 Jones K.S. A statistical interpretation of term specificity and its application in retrieval // Journal of Documentation: журнал. – MCB University: MCB University Press, 2004. – Vol. 60, №5. – P. 493-502.

34 Kureychik V., Semenova A. Combined Method for Integration of Heterogeneous Ontology Models for Big Data Processing and Analysis / in: Silhavy R., Senkerik R., Kominkova Oplatkova Z., Prokopova Z., Silhavy P. (eds) Artificial Intelligence Trends in Intelligent Systems. CSOC 2017 // Advances in Intelligent Systems and Computing. – Springer; Cham, 2017. – Vol. 573. - P. 302-311.

35 Gupta S., Szekely P., Knoblock A.C., Goel A., Taheriyan M., Muslea M. Karma: A System for Mapping Structured Sources into the Semantic Web // Extended Semantic Web Conference. ESWC. – Springer, 2015. – P. 430-434.

36 Semenova A.V., Kureychik V.M. Multi-objective particle swarm optimization for ontology alignment // 10th International Conference on Application of Information and Communication Technologies. - AICT, 2016. - 159 p.

37 Cochez M., Ristoski P., Ponzetto S.P., Paulheim H. Biased graph walks for rdf graph embeddings // Proceedings of the 7th International Conference on Web Intelligence. ACM. – Mining; Semantics, 2017. - P. 21.

38 Утепбергенов И.Т., А.И.Буранбаева, А.А.Иванов, А.Н.Нургулжанова. Қазақстанда инновациялық қызметтің өмірлік циклінің кезеңдерін ақпараттық қамтамасыз етуді құрудың әдістемелік принциптері // Вестник КазАТК (Спец. выпуск) Том 2, Алматы, 2019 г.

39 И.Т. Утепбергенов, Д.Т. Касымова, Д.М. Ескендирова, А.Т. Ахмедиярова Подход к выявлению и устранению семантических противоречий в «больших данных» // Вестник КазАТК №2-2017 ISSN 1609-1817 с. 200-206

40 Шереметьева С. О. Linguistic models and tools for processing patent claims: monograph // Russian Federation, Ministry of education and science, South Ural state university, Department of linguistics and international relations. – Chelyabinsk: SUSU publishing centre, 2017. – 156 p.

41 Семерханов И.А. Методы и алгоритмы автоматизированной интеграции информационных ресурсов на основе онтологического подхода: дис. ... канд. техн. наук: 05.13.12. – Спб., 2014. – 140 с

42 Левенштейн В.И. Двоичные коды с исправлением выпадений, вставок и замещений символов // ДАН СССР. - 1965. - Т. 163, №4. - С. 845–848.

43 Afonin Sergey, Golomazov Denis. Calculating Semantic Similarity Between Facts // Proceedings of the International Conference on Knowledge Discovery and

Information Retrieval (KDIR) / ed. by Ana Fred. Valencia. - Spain: INSTICC Press, 2010. - P. 514–517.

44 Семенова А.В., Курейчик В.М. Ансамбль классификаторов для автоматического пополнения онтологий // Известия ЮФУ. Технические науки. - 2018. - №2 (196). – С. 163-137.

45 Uskenbayeva R., Temirbolatova T., Young Im Cho, Uskenbayeva Z., Bektemyssova G., Kassymova A. Recursive decomposition as a method for integrating heterogeneous data sources // Proceedings of the 15th International Conference on Control, Automation and Systems (ICCAS). – Busan; South Korea, 2015. – P. 2076-2079.

46 Ускенбаев Р.К., Темірболатова Т.Т., Касымова А.Б. Бұлттық есептеуде mapreduce технологиясымен үлкен деректерді өңдеу // Вестник КазНТУ имени К.Сатпаева. – 2015. - №5(111). - С. 50-53.

47 Ускенбаева Р.К., Аманжолова С.Т., Темірболатова Т.Т. Анализ и локализация инцидентов снижения работоспособности распределенных вычислительных систем // Труды международного форума «инженерное образование и наука в XXI веке: проблемы и перспективы», посвященного 80-летию Каз НТУ им. К.И. Сатпаева. – Алматы, 2012. – С. 354-361.

48 Chinibayeva T. Security semantic database problems // Herald of the Kazakh-british technical university // V International Conference "Digital technology in science and industry - 2019» (dtsi-2019): 10th anniversary information technology international university. – 2019. - Vol. 16, №3. - P. 168-174.

49 Uskenbayeva R., Chinibayeva T. Algorithm for the construction of an ontology in the field of scientific knowledge // The Bulletin of Kazakh Academy of Transport and Communications named after M. Tynyshpayev. – 2018. - Vol. 107, №4. - P. 259-266.

50 Uskenbayeva R., Chinibayeva T. Method of extracting meta description from databases // Herald of the Kazakh-british technical university. - 2018. - Vol. 15, №4. – P. 116-123.

51 Uskenbayeva R., Chinibayeva T. Model, data integration algorithms of information systems based on ontology // Journal of Theoretical and Applied Information Technology. - 2021. - Vol. 99, №9. – P. 2125-2143.

52 Temirbolatova T., Jarmukhambetov Y., Temirbolatova U. The method of extracting semantic meta descriptions from databases // 2nd International scientific conference «Information Technologies in Science &Industry» International IT University. - Almaty, 2016, may 19. – P. 115-117.

53 Temirbolatova T., Beisenov D. Automatic asynchronous exchange of business object between heterogeneous systems // The 12th ICIT&M 2014. - Riga; Latvia: Information Systems Management Institute, 2014. – 273 p.

54 Temirbolatova T., Khamitov A., Keldybay A., Sembayeva T. Manage different-structured Big Data // The 12th ICIT&M 2014. – Riga; Latvia: Information Systems Management Institute, 2014. – 756 p.

55 Ускенбаева Р., Бектемысова Г., Темірболатова Т. Интеграция больших неоднородных данных с использованием языка R и HADOOP // Вестник КазАТК. – 2015. - №4. – С. 44-49.

56 Uskenbayeva R., Chinibayev Y., Kassymova A., Temirbolatova T., Mukhanov K. Technology of integration of diverse databases on the example of medical records // Proceedings of the 14th International Conference on Control, Automation and Systems (ICCAS). - Gyeonggi–do; Korea, 2014. – P. 282-285.

57 Ускенбаева Р.К., Темірболатова Т.Т., Курмангалиева Б.К. Применение методов Process Mining для оптимизации моделей бизнес-процессов на примере информационной системы выдачи лицензий и разрешительных документов e-license: практическое исследование // Вестник КазНТУ. – 2015. - №3. – С. 50-55.

58 Uskenbaeva R., Temirbolatova T., Bektemyssova G., Chinibaev Y. Recovery and visualization of integrated database // The 12th ICIT&M 2014. - Riga; Latvia: Information Systems Management Institute, 2014. –258 p.

59 Uskenbayeva R., Amanzholova S., Khasenova G.I., Temirbolatova T. Analysis and Localization of Performance Degradation Incidents of Distributed Computer Systems // Smart-government: Proceeding of the International Scientific-Practical Conference. – Астана, 2014. - P. 75-81.

60 Temirbolatova T., Chinibaev Y. Development of the augmented reality applications based on ontologies // Computer modelling & New technologies. – Riga; Latvia, 2016. - Vol. 20, №4. – P. 18-22.

61 Uskenbayeva R., Kurmangaliyeva B., Temirbolatova T. Application Process Mining techniques to optimize the business process models based on Information systems issuing licenses and permits e-license: practical research // SICE. – Hangzhou; Chine, 2015. – P. 298-300.

62 Han H., Zha H., Giles C.L. Name disambiguation in author citations using a K-way spectral clustering method // Digital Libraries - 2005. JCDL'05. Proceedings of the 5th ACM/IEEE-CS Joint Conference. - IEEE, 2007. - P. 334–343.

63 Byron W., Thomas T. Semi-automated screening of biomedical citations for systematic reviews // BMC Bioinformatics. - 2010. - Vol. 11, №1. - P. 55.

64 McCallum A., Nigam K., Ungar L.H. Efficient clustering of high-dimensional data sets with application to reference matching // Proc. of the 6th ACM SIGKDD Int. Conf. on Knowledge discovery and data mining. - ACM, 2000. - P. 169–178.

65 Мухамедиев Р. И., Мусабаев Р. Р., Булдыбаев Т., Кучин Я., Сымагулов А., Оспанова У., Якунин К., Мурзахметов С., Сагындык Б. Эксперименты по оценке средств массовой информации на основе тематической модели корпуса текстов. Cloud of Science. 2020. Т. 7. No 1, С. 87-101 (Входит в перечень ВАК РФ, Импакт-фактор РИНЦ – 0.482).

66 Бектемысова Г.У., Байер Д., Мукажанов Н., Едильхан Д., Касымова А.Б., Оразбеков С. Масштабируемые решения для управления большим объемом данных // Вестник КазНТУ имени К.Сатпаева. – 2015. - №5 (111). - С. 111-118.

67 Yakunin K., Kuchin Y., Mukhamediev R., Musabayev R. Classification of negative publication in mass media using topic modeling// Journal of Physics: Conf. Series. – 2020. – С. 1-12 (Scopus, Cite Score =0.7)

68 Дорофеева В.И., Никольский Д.Н., Федяев Ю.С. Организация электронного документо оборота при планировании научно-исследовательской работы кафедры вуза // Электронное информационное пространство для науки,

образования, культуры: материалы III Международной научно-практической интернет-конференции. – Орел: ООО «Горизонт», 2015. – С. 129-135.

69 Дорофеева В.И., Мотин А.Г., Никольский Д.Н., Федяев Ю.С. Разработка веб-приложения для автоматической генерации отчета о научно-исследовательской работе структурного подразделения вуза // Современные методы прикладной математики, теории управления и компьютерных технологий: сб. тр. VIII междунар. конф. «ПМТУКТ-2015». – Воронеж: Изд-во «Научная книга», 2015. – С. 124-127.

70 Загорулько Ю.А., Боровикова О.И., Загорулько Г.Б. Применение паттернов онтологического проектирования при разработке онтологий научных предметных областей // Труды XIX Международной конференции «Аналитика и управление данными в областях с интенсивным использованием данных». – М., 2017. – С. 258 -265.

71 Бубарева О.А. Интеграция неоднородных онтологий на основе их семантической близости // Информация и образование: границы коммуникаций INFO'12: сборник научных трудов. – ГорноАлтайск: РИО ГАГУ, 2012. - №4(12). – С. 456−458.

72 Глушков Н. А. Анализ методов тематического моделирования текстов на естественном языке // Молодой ученый. – 2018. – №19. – С. 101-103.

73 Нугуманова А. Б. Предметно-ориентированные модели и методы распределенного поиска, обработки и анализа текстовой информации в сети Интернет [Electronic resource] : диссертация на соискание ученой степени доктора философии : 6D07030 "Информационные системы" / Нугуманова Алия Багдатовна; Вост.-Каз. гос. техн. ун-т им. Д. Серикбаева - Усть-Каменогорск : [б. и.], 2014 . - Электрон. текст. дан., 1 электр. опт. диск (CD-ROM) . - Загл. с контейнера Библиогр.: с. 118-125 - [б. т.].

74 Самбетбаева М.А. Ғылыми-білім беру қызметін қолдауға арналған ақпараттық технологиялар бойынша қазақ тілі морфологиясын ескере отырып көптілді тезаурус жасау [Электрондық ресурстар] : философия докторы (PhD) дәрежесін алу үшін дайындалған диссертация : 6D070300 – Ақпараттық жүйелер / М. А. Самбетбаева,; ғыл. кеңесшілер: Ж. А. Тусупов, А. М. Федотов ; Л. Н. Гумилев атын. Еуразия ұлттық ун-ті - Астана : [б. ж.], 2016 . - Электрон. мәлімет , 1 электрон. опт. диск (CD-ROM) . Библиогр.: 140-148 б. - [т. ж.] .

75 Sharipbay A.A., Zhetkenbay L., Bekmanova G., Kamanur U. The ontological model of noun for Kazakh-turkish machine translation system. Proceedings of the international conference "Turkic languages processing", TurkLang 2015, September 17-19, 2015, Kazan, Tatarstan, Russia, ISBN 978-59690-0262-3, -P. 15.

76 Sharipbay A.A., Zhetkenbay L., Bekmanova G., Kamanur U. The ontological model of noun for Kazakh-turkish machine translation system. Proceedings of the international conference "Turkic languages processing", TurkLang 2015, September 17-19, 2015, Kazan, Tatarstan, Russia, ISBN 978-59690-0262-3, -P. 15.

77 Найханова Л.В. Методы и модели автоматического построения онтологий на основе генетического и автоматного программирования: автореф. ... док. техн. наук: 05.13.11. – М., 2009. - 35 с.

ҚОСЫМША А

ИП «TRANS.KZ»

Almatycity, Taugul 3 district, Surtibayeva, 736
БИН 631105300143, ИИК KZ74914002204US02KX0, ДБ АО «СБЕРБАНК» г.Алматы
SABRKZKA

Акт о внедрении
результатов диссертационной работы Чинибаевой Толганай Темірболаткызы
представленной на соискание степени доктора философии(PhD)
по специальности: 6D070400 – «Вычислительная техника
и программное обеспечение»

Настоящим подтверждаем, что результаты диссертационного исследования Чинибаевой Толганай на тему «Модели и методы управления гетерогенными данными (BigData)» обладают актуальностью, представляют практический интерес и были использованы в виде программного модуля, как новый подход по взаимодействию совместного использования информации и координированного принятия решений. Формальная онтология управления цепью поставок повысила возможность взаимодействия различных звеньев цепи поставок.

Использование онтологии для проектирования и создания информационной системы управления цепью поставок привело к созданию легко интегрируемой и многократно используемой базы знаний.

ИП «Trans.kz» выражает глубокую признательность Чинибаевой Толганай за предоставленную возможность практического применения столь интересных результатов диссертационного исследования и надеется на активное продолжение ее работ и нашего сотрудничества.

Директор _____ Кайбаров Марат

ИП «Trans.kz»

ҚОСЫМША Б

Оңтүстік Корея Сеул қаласындағы Гачон университетінде тағылымда өтілгендігі туралы сертификат

가천대학교
Gachon University

Certificate of Achievement

This certifies that, PhD Candidate of The International Information Technology University, Almaty, Kazakhstan

Temirbolatova Tolganay

has successfully completed the scientific internship at Gachon University (Seoul, South Korea) from December 14, 2015 to December 24, 2015

Supervisor Professor Young Im Cho

Department of Computer Engineering

ҚОСЫМША В

Бағдарлама листингі

```
package edu.ucla.sspace.doc.reader;

import java.io.BufferedReader;
import java.io.File;
import java.io.FileInputStream;
import java.util.List;
import java.util.Map;

import org.apache.poi.hwpf.HWPFDocument;
import org.apache.poi.hwpf.extractor.WordExtractor;
import org.apache.poi.xwpf.usermodel.XWPFDocument;
import org.apache.poi.xwpf.usermodel.XWPFParagraph;
import com.orientechnologies.orient.core.db.document.ODatabaseDocumentTx;
import com.orientechnologies.orient.core.intent.OIntentMassiveInsert;
import com.orientechnologies.orient.core.record.impl.ODocument;
import com.orientechnologies.orient.core.sql.query.OSQLSynchQuery;

import edu.ucla.sspace.basis.StringBasisMapping;
import edu.ucla.sspace.lsa.LatentSemanticAnalysis;
import edu.ucla.sspace.matrix.NoTransform;
import edu.ucla.sspace.matrix.factorization.SingularValueDecompositionLibC;

import java.io.FileFilter;
import java.io.IOException;
import java.io.StringReader;

import org.apache.commons.io.FileUtils;
import java.util.ArrayList;
import java.util.HashMap;

public class DocReader {
  public static void readDocxFile(String fileName) throws IOException {
        int countDocx = 0;
        File folder = new File(fileName);
        File[] listOfFiles = folder.listFiles();
        Map<String, String> listofDoc = new HashMap<>();
        for (int i = 0; i < listOfFiles.length; i++) {
                countDocx++;
                File file = listOfFiles[i];

                if (file.isFile() && file.getName().endsWith(".docx")) {
```

```
                String content = FileUtils.readFileToString(file);
                try {
                        FileInputStream fis = new
FileInputStream(file.getAbsolutePath());
                        XWPFDocument document = new XWPFDocument(fis);
                        List<XWPFParagraph> paragraphs =
document.getParagraphs();
                        StringBuilder pars = new StringBuilder();
                        for (XWPFParagraph para : paragraphs) {
                                pars.append(para.getText() + "\n");
                        }
                        listofDoc.put(file.getName(), pars.toString());
                        fis.close();
                } catch (Exception e) {
                        System.err.println(file.getName());
//                      e.printStackTrace();
                }
        }
        if (file.isFile() && file.getName().endsWith(".doc")) {
                try {
                        FileInputStream fis = new
FileInputStream(file.getAbsolutePath());
                        HWPFDocument doc = new HWPFDocument(fis);
                        WordExtractor we = new WordExtractor(doc);
                        String[] paragraphs = we.getParagraphText();
                        StringBuilder pars = new StringBuilder();
                        for (String para : paragraphs) {
                                pars.append(para.toString() + "\n");
                        }
                        listofDoc.put(file.getName(),pars.toString());
                        fis.close();
                } catch (Exception e) {
                        System.err.println(file.getName());
//                      e.printStackTrace();
                }
        }
    }
    LatentSemanticAnalysis lsa = new LatentSemanticAnalysis(true, 2, new
NoTransform(),
                new SingularValueDecompositionLibC(), false, new
StringBasisMapping());
    listofDoc.forEach((key, value) -> {
        try {
                lsa.processDocument(new BufferedReader(new
StringReader(value)));
```

```
                } catch (IOException e) {
                        // TODO Auto-generated catch block
                        e.printStackTrace();
                }
        });


        lsa.processSpace(System.getProperties());
    }

    public static void main(String[] args) throws IOException {

        readDocxFile("C:\\Users\\tolganay\\Desktop\\data");

        // readDocFile("C:\\Auction\\pdfReader\\src\\pdfreader\\test3.doc");

        // ODatabaseDocumentTx db = new ODatabaseDocumentTx("local:test");
        // db.open("admin","admin"); // db.create();

        // ODocument doc = new ODocument("Person");
        // doc.field("name", "Luke");
        // doc.field("surname", "Skywalker");
        // doc.field("city", new ODocument("City").field("name",
        // "Rome").field("country", "Italy"));
        // doc.save();

        // for (ODocument animal : db.browseClass("Person")) {
        // System.out.println( animal.field( "name") + " " + animal.field(
        // "surname"));
        //
        // }
        // db.close();

    }
}
```