Т.Т. Чинибаеваның 6D070400 – Есептеу техникасы және бағдарламалық қамтамасыз ету мамандығы бойынша философия докторы (PhD) дәрежесін алуға ұсынылған «Гетерогенді құрылымы бар деректерді басқаруға арналған үлгілер мен әдістер (Big Data)» тақырыбындағы диссертациялық жұмысына

АНДАТПА

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 байланысты.

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

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 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 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 [1-12], 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 деректер технологиясын заманауи жай-күйі мен нарықта алатын орнын көңіл бөлінді. 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 келтірілген.

Кесте 1 – 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» жүйесі, Астрахан мемлекеттік университетінің «Ғылыми қызмет нәтижелері» ақпараттық -аналитикалық жүйесі, Elsevier-дің PURE жүйесі ұйымның ғылыми қызметі мен өнімділігін бағалауға кеңінен мониторинг жүргізеді. Әлемдегі ғылыми ақпараттарды басқаруда қолданылатын ірі веб қызметтерінің сипаттамаларының салыстырмасы кесте 2 берілген.

Бірінші тараудың қорытындысында ғылыми деректерді өңдеу мен талдаудың қолданыстағы жүйесі туралы белгілі ақпараттың негізгі кемшіліктері тізіліп, оларды негізгі мәселенің мүмкін шешімдері ретінде қарастыруға болады. Бұл кемшіліктерге мыналар жатады: мәліметтерді енгізудің күрделілігі; ақпаратты іздеудің күрделілігі және төмен мүмкіндігі; қатаң және ақпаратсыз білім үлгілерін пайдалану, жүйелік икемділіктің болмауы; пайдаланушының жартылай автоматты енгізуі үшін емес, интернеттен ақпаратты өңдеу бағыты; ақпаратты жүктеу, өңдеу және алу алгоритмдерін интеллектуалдандыруға жеткіліксіз көңіл бөлу.

Кесте 2 – 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
Ш	1	Bibster	Карлсруэ университеті,	Жүйеден деректерді RDF	Жүйеге ақпарат құрылымдалған	BibTeX

			Амстердам университеті, Dresden Bank	форматында шығарады	файл арқылы жүктеледі	
	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

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

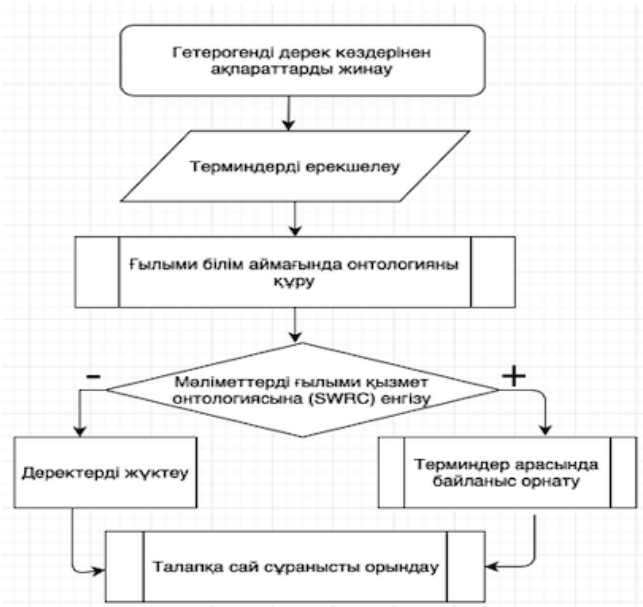
Автоматтандырылған жүйенің негізгі прототипі болып табылатын ғылыми-техникалық ақпараттың қорытындысы мен күрделі ұймдастырған есептеу үрдісінің жалпы формальды үлгісі.

D ғылыми білім аймағы берілді деп есептейік (мысалы, Computer science). I – білімнің осы аймағының шеңберіндегі ғылыми-техникалық ақпараттың бірліктерін сипаттау жиыны болсын (атомдық өлшем). Бұндай бірліктерге мыналар жатады: ғылыми мақалалар; патенттер; есептер; конференцияларда оқылатын есеп; есепке арналған тезистер; монографиялар; білім беруге арналған көмекші құрал мен басқа авторлық туындылар (рефераттар, аудармалар). I жиынының әрбір элементі сәйкес келетін объектінің кейбір мәтіндік сипаттамасынан тұрады.

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

Жүйенің жалпы сызбасы сурет 1 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 білімнің аймағындағы қарастырылатын онтологияны құру;
- ғылыми негізінің деректерін жүктеу;
- Білім аймағында құрастырылған экземпляр мен ғылыми салалардың нәтижелері жайлы жүктелген ақпараттардың арасындағы байланысты орнату;
- Алынған ақпаратқа аналитикалық сұранысты орындау;
- D ғылыми білім аймағын сипаттайтын терминдерді ерекшелеп алу (кілт сөз);
- D ғылыми білім аймағының онтологиясын құрастыру;
- Ақпараттардан алынатын деректерді жүктеу ғылыми саламен ерекшеленеді;
- Құрастырылған онтология түсінігі мен қолданушылардың ғылыми қорытындыларының арасындағы байланысты орнату;
- Сұраныстарды орындайтын үлгі алынған ақпараттың қорытындысын қамтамасыз етеді.



Сурет 1 – Ғылыми ақпаратты басқару жүйесінің жалпы архитектурасы

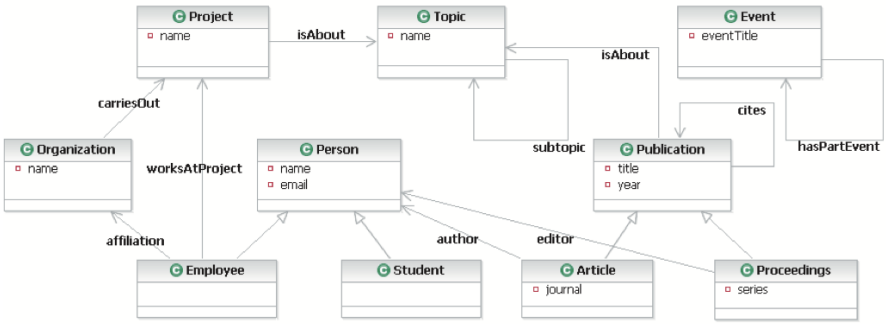
Білім аймағын сипаттайтын терминдерді ерекшелеуде семантикалық әдіс қолданылды.

Деректерді енгізудің келесі тәсілдері жоспарланады:

библиографиялық сілтемелерді реттеу;

жүктелетін мета деректерді реттеу (BibTeX, MathML, LaTeX, FinXML);

жолдарды қолмен толтыру.



Сурет 2 – SWRC онтологиясы (фрагмент)

Библиографиялық сілтемелерді реттеу. Библиографиялық сілтемелерден ақпаратты алу ақпаратты құрамы жасалынбаған мәтіннен алу тапсырма болып табылады. Берілген библиографиялық сілтемелердің реттелу әдісін тестілеудің нәтиже көрсеткіші бойынша аса жоғарғы әсерін көрсеткен Conditional Random Fields (CRF) алгоритмі. АҚШ Браун университетінде жасалынған FreeCite бағдарламалық комплекс осы алгоритмді іске асыратын CRF++ кітапханасында қолданды.

```
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

“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”.

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

Білім аймағындағы қалыптасқан үлгі мен ғалымдардың ғылыми еңбектерінің нәтижесінен тұратын жүктелген мәтіндерден алынған ақпараттардың арасындағы байланыстың орнауы аналитикалық сұраныстарды орындау үшін аса қажет. Бұл деңгейге дейін қолданылатын құжаттан қызметкердің ғылыми еңбегі жайлы мөлшері бар ақпарат қана ерекшеленіп алынды.

Білімге байланысты онтологиялық әрекет аналитикалық сұраныстарды орындайтын алгоритмдердің қазіргі және бұрынғы апробациясын қолдануға мүмкіндік береді. Жекелей алғанда, онтологияны қолдану арқылы сұранысты қайта жазу логикалық қорытындының механизмінің көмегі арқылы автоматты түрде орындау мүмкін.

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.
}
```

Берілген тематикалық бөлімдермен бірге мәтіндер жиынтығынан терминдерді іріктеп алу алгоритмінің сипаты. $W - \varepsilon$ – бос сөзін қоскқанда Doc коллекциялары арқылы беріліп, құжаттардың барлығында кездесетін барша сөздердің жиыны болсын делік, ал PW – барлық реттелінген сөздердің жұптары 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:

- $\tau_1: PW \to W, \tau_2: PW \to W$ – жұптар сөзі 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 көпше түріндегі f функциясының орташа мәні.

Мысалы, $Av(|\cdot|, R)$ – айдардағы құжаттардың орташа саны, $Av(L, Doc)$ - құжаттың орташа ұзындығы, $Av(id, A)$- $A = \{a_1, \dots, a_k\}$ жиынтығының орташа арифметикалық саны.

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.

Ғылыми білімнің жеке аймағындағы онтологияны құрастырудың алгоритмі ғылыми конференциялардың хабарландыруларының жинағына,

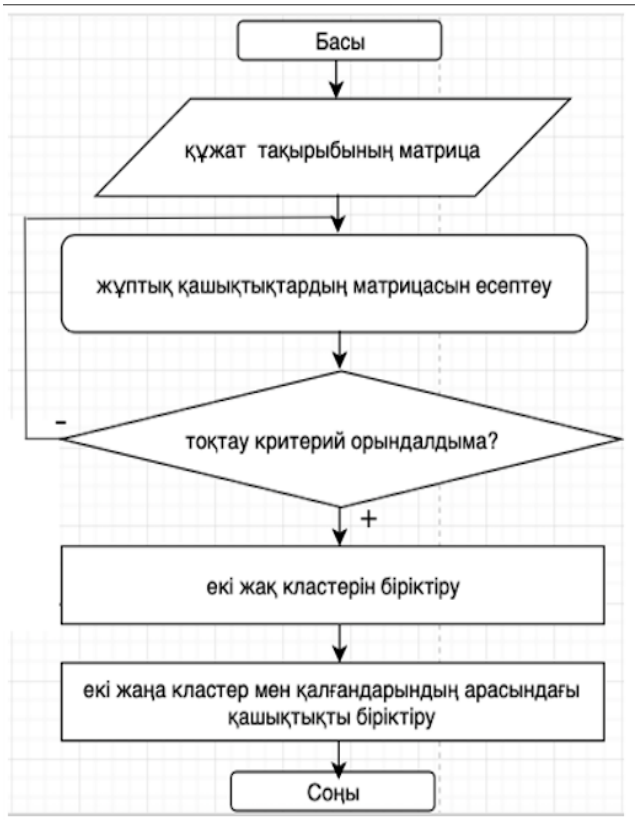
тақырыптарға бөлінген автордың ғылыми білім саласының онтологиясын құруға арналған, сондай-ақ Интернеттегі іздеу жүйесінен алынған ақпарат берілген. Контактілерге арналған хабарламалар (CFP) деп аталады, онтологияны құру үшін негізгі деректер көзі ретінде қолданылды.

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:

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

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

$$A\ is * \ keyword * prep(aux)?\ B$$



Сурет 3 - ғылыми білімнің жеке аймағындағы онтологияны құрастырудың алгоритмдік сызбасы

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.

Жұмыстың аппробациясы:

1. R. Uskenbayeva, Y. Chinibayev, A. Kassymova, T. Temirbolatova, K. Mukhanov. Technology of integration of diverse databases on the example of medical records//Proceedings of the 14th International Conference on Control, Automation and Systems (ICCAS 2014) - Gyeonggi -do, Korea, 2014. P 282-285. ISSN: 2093- 7121.
2. R.Uskenbayeva, T.Temirbolatova, Young Im Cho, Z.Uskenbayeva, G.Bektemyssova, A. Kassymova. Recursive decomposition as a method for integrating heterogeneous data sources//Proceedings of the 15th International Conference on Control, Automation and Systems (ICCAS 2015). – Busan, South Korea. October 13-16, 2015 – P.2076-2079. ISSN: 2093 - 7121
3. Р.К. Ускенбаев, Т.Т.Темірболатова, А.Б. Касымова. Бұлттық есептеуде mapreduce технологиясымен үлкен деректерді өңдеу - // Вестник КазНТУ имени К.Сатпаева No5 (111). – 2015. С.50 - 53. ISSN 1680- 9211
4. Р.Ускенбаева, Г. Бектемысова, Т.Темірболатова. Интеграция больших неоднородных данных с использованием языка R и HADOOP - Вестник КазАТК - №4 2015-11-01
5. 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
6. T. Temirbolatova, D. Beisenov Automatic asynchronous exchange of business object between heterogeneous systems - The 12th ICIT&M 2014. 2014 April 16-17, 2014, Information Systems Management Institute, Riga, Latvia
7. T. Temirbolatova, A.Khamitov, A. Keldybay, T.Sembayeva Manage different-structured Big Data - The 12th ICIT&M 2014. 2014 April 16-17, 2014, Information Systems Management Institute, Riga, Latvia
8. 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, May 19, 2016 Almaty, Kazakhstan. ISBN 978-601-7407-33-9
9. T. Chinibayeva Security semantic database problems // Herald of the Kazakh-british technical university ISSN1998-6688. V INTERNATIONAL CONFERENCE "DIGITAL TECHNOLOGY IN SCIENCE AND INDUSTRY - 2019» (DTSI-2019), 10th Anniversary INFORMATION TECHNOLOGY INTERNATIONAL UNIVERSITY Vol.16, No.3 (2019), pp. 168-174

10. R.Uskenbayeva, T.Chinibayeva. 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 ISSN 1609-1817. Vol. 107, No.4 (2018), pp. 259-266

11. R.Uskenbayeva, T.Chinibayeva. Method of extracting meta description from databases//Herald of the Kazakh-british technical university ISSN1998-6688. Vol.15, No.4 (2018), pp. 116-123

12. 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