## *Раздел 2. «Информационно-коммуникационные технологии»*

УДК 004.021
МРНТИ 20.15.05

Садуакасов А.А., Мухаметжанова Б.О.

*Әбілқас Сағынов атындағы Қарағанды техникалық университеті,*
*Қарағанды, Қазақстан*
*(E-mail.ru: arkhatsaduakasov@gmail.com, grek79@mail.ru )*

# Семантикалық іздеу алгоритмдерін жүйелі және салыстырмалы талдау

Бүгінгі цифрлық дәуірде пайдаланушылар күнделікті оңай қол жетімді және үнемі жасалатын деректердің үлкен көлеміне байланысты желіде қажетті ақпаратты табу қиынға соғады. Мәселе іздеу процесін қиындататын деректердің үлкен көлемінен туындайды және дәстүрлі кілт сөзге негізделген іздеу жүйелері күрделі сұрауларды өңдеуде тиімсіз болуы мүмкін. Нәтижесінде пайдаланушылар маңызды емес немесе толық емес іздеу нәтижелеріне тап болуы мүмкін, бұл оларға қажетті ақпаратты табуды және алуды қиындатады.Бұл мәселені машиналық оқыту мен табиғи тілді өңдеудің озық әдістеріне негізделген семантикалық іздеуді қолдану арқылы сәтті шешуге болады. Семантикалық іздеу жүйелерге пайдаланушы сұрауларының мағынасы мен контекстін тереңірек түсіндіруге мүмкіндік беретін инновациялық тәсіл болып табылады. Бұл әдіс жеке кілт сөздерді есепке алып қана қоймай, олардың семантикалық байланыстары мен контекстін талдауға мүмкіндік береді, бұл сайып келгенде дәлірек және сәйкес іздеу нәтижелерін береді. Осылайша, семантикалық іздеу пайдаланушыларға онлайн кеңістікте ақпаратты іздеуде тиімдірек және қанағаттанарлық тәжірибені қамтамасыз ететін негізгі құралға айналады.

*Кілт сөздер:* табиғи тілді өңдеу, семантикалық іздеу, ақпаратты алу, ақпаратты іздеу, үлкен деректер көлемі, дәстүрлі іздеу жүйелері, сөз векторлары, модель өнімділігі.

## *Кіріспе*

Қазіргі цифрлық дәуірде Интернеттегі деректердің үлкен көлемі пайдаланушыларды қажетті ақпаратқа тез және тиімді қол жеткізуге шақырады. Дәстүрлі кілт сөзге негізделген іздеу жүйелерін пайдалану қанағаттанарлықсыз болуы мүмкін, өйткені олар жиі күрделі сұрауларды өңдей алмайды, нәтижесінде маңызды емес немесе толық емес нәтижелер қайтарылады. Семантикалық іздеу пайдаланушы сұрауларын тереңірек түсіну үшін машиналық оқытуды және табиғи тілді өңдеуді қолдана отырып, осы мәселеге жауап болып табылады.

Семантикалық іздеу дәлірек және сәйкес нәтижелер беру арқылы пайдаланушы тәжірибесін жақсартудың кілті болып табылады. Бұған қоса, бұл тәсіл компаниялар мен ұйымдарға деректер мен білім ресурстарын тиімдірек басқаруға мүмкіндік беріп, анағұрлым негізделген шешімдер қабылдауға және жалпы тиімділікті арттыруға мүмкіндік береді.

Табиғи тілді өңдеу және машиналық оқыту әдістерін пайдалана отырып, семантикалық іздеу пайдаланушы сұрауларын дәлірек түсіндіруге және веб-беттердің мазмұнын талдауға мүмкіндік береді. Семантикалық іздеу жүйелері сұраудың мәнмәтінін және мақсатын түсінуге тырысады, нәтижесінде дәстүрлі кілт сөздерді іздеуге қарағанда дәлірек іздеу нәтижелері алынады [1].

Семантикалық іздеуде табиғи тілдегі сұрауларды іздеу мүмкіндігі де бар, бұл дәстүрлі кілт сөзге негізделген әдістерге қарағанда маңызды артықшылық болып табылады. Ыңғайлылықты арттырудан басқа, бұл стандартты әдістерді қолдану арқылы жіберіп алуы мүмкін ақпаратты толық және кеңірек қамтуды қамтамасыз етеді.

Күнделікті жұмыста семантикалық іздеу нәтижелердің дәлдігі мен өзектілігін арттыруға көмектеседі. Кілт сөздерді жай ғана сәйкестендіретін дәстүрлі іздеу жүйелерінен айырмашылығы, семантикалық іздеу сұраудың мағынасын және беттердің мазмұнын түсінуге тырысады, нәтижесінде дәлірек нәтижелер алынады. Бұл әдіс ақпаратты іздеуді жақсартады, әсіресе бір нысанды сипаттау үшін синонимдер немесе әртүрлі тұжырымдар қолданылуы мүмкін жағдайларда [2].

Семантикалық іздеу күнделікті тапсырмаларда маңызды артықшылықтар береді, бұл пайдаланушыларға қажетті ақпаратты тезірек және оңай табуға көмектеседі. Бұл тәсіл нәтижелердің өзектілігі мен дәлдігін жақсартуға көмектеседі, бұл қазіргі ақпаратқа бай әлемде тиімді іздеуді қамтамасыз етудің негізгі факторы болып табылады.

*Негізгі бөлім*

Соңғы жылдары табиғи тілді өңдеуге көп көңіл бөлініп, ауқымды зерттеулердің нысаны болды. Бұл салада семантикалық іздеу мәселелерін шешуге мамандандырылған көптеген модельдер ұсынылды. Платформаның екі маңызды құрамдас бөлігі талқыланады: кеңейтілетін және икемді құрылымды қамтамасыз ететін онтологияны оқыту және нысанды біріктіру. Бұл құрылым семантикалық іздеумен байланысты жалпы тапсырмалар мен мәселелерді шешеді.

Сонымен қатар, зерттеу бағаланған семантикалық іздеу әдістемелері мен қозғалтқыштарымен проблемаларды талдау және анықтау үшін төрт перспективаны пайдаланады. Бұл зерттеу ақпаратты іздеу тапсырмасында қолданылатын әртүрлі рейтингтік функциялардың егжей-тегжейлі талдауын ұсынады. Іздеу процесінде мүмкіндік жиілігі, кері құжат жиілігі және ұзындықты қалыпқа келтіру белсенді қолданылады. Сұрау мен құжаттың ұқсастық дәрежесі, жалпы сөз тіркестерінің әсері бағаланады [3].

Сондай-ақ мақалада термин жиілігін, кері құжат жиілігін және келесі сөзді терістеуді біріктіретін мәтіндік сезімді санаттау әдісі бар. Мәтінді жіктеу сөздердің екілік қапшығының нәтижелерін, мәтіндік сезімді санаттау және мәтіндік сезімді санаттау үлгілерін келесі сөзді терістеу үлгісімен салыстырады. Екі жаңа үлгі архитектурасы енгізілді: үздіксіз сөздер қаптамасының үлгісі және үздіксіз скипграмма моделі, бай деректерден сөздердің векторлық көрсетілімдерін жасауға және бар әдістермен салыстыруға бағытталған [4].

Сондай-ақ мақалада қайталанулар мен қисықтарды болдырмайтын зейін механизмдеріне негізделген «Трансформер» деп аталатын жаңа желі архитектурасы талқыланады. Бұл архитектураның жоғары сапасын және алдыңғы әдістермен салыстырғанда жоғары параллельді және төмен оқу уақытын көрсететін екі машиналық аударма тапсырмасы ұсынылған. Ұсынылған үлгілер төмен есептеу шығындарымен дәлдікте айтарлықтай жақсартуларды қамтамасыз етеді.

Сондай-ақ мақалада алдын ала дайындалған сөйлем деңгейіндегі ендірулер және «Трансформер» үлгісіндегі екі жақты кодтаушы көріністерді біріктіретін тілді көрсету үлгісі талқыланады. «Трансформер» деп аталатын бұл модель табиғи тілді өңдеудің он бір тапсырмасын жақсы орындайды. Сонымен қатар, сөйлемді кодтаудың екі моделі ұсынылады: «Трансформер»-ге және терең орташалау желісіне негізделген. Бұл модельдер дәлдік пен есептеу ресурстары арасындағы келісуді ұсынады.

Зерттеу сонымен қатар екі жақты кодтаушы көрсетілімдерін біріктіретін және екі жаңа архитектураны енгізетін тілді ұсыну үлгісін ұсынады: үздіксіз сөздер қаптамасы моделі және үздіксіз өткізіп жіберу үлгісі. Бұл архитектуралар есептеу шығындарын азайта отырып, дәлдікті айтарлықтай жақсартады.

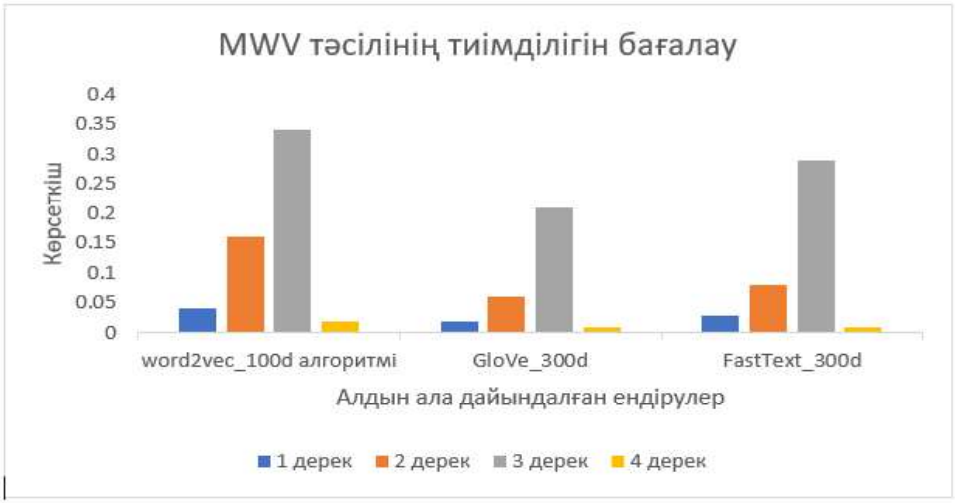
Деректер жинағы CISI деректер жинағы, Глазго университеті жариялаған мәтінге негізделген деректер жинағы ақпаратты іздеу үшін пайдаланылады. Сұрау-құжат сәйкес келетін негізгі шындық CISI.REL файлында қамтылған, ол оқытылған үлгілерді салыстыру және олардың өнімділігін бағалау үшін пайдаланылуы мүмкін [5].

Машиналық оқыту үлгілерінің дәлдігі мен сенімділігін арттыру үшін деректерді дайындау өте маңызды. Ол өңделмеген мәтіндік деректерді тазалауды және өзгертуді талап етеді, бұл деректерден маңызды қорытындылар мен үлгілерді шығаруды жеңілдетеді. Алдын ала өңдеу сонымен қатар табиғи тілді өңдеу алгоритмдерінің тиімділігін арттыратын және есептеу уақытын қысқартатын деректер көлемін азайтуға көмектеседі. Деректерді дайындау машиналық оқыту моделінің сапасы мен тиімділігіне үлкен әсер етеді. Төменде тізімделген деректерді алдын ала өңдеу қадамдары деректерді машиналық оқыту үлгілеріне бермес бұрын дайындау үшін пайдаланылады. Мәтін мәтінді тазалау қадамында тазартылады, ол сонымен қатар барлық мәтінді кіші әріптермен және барлық таңбалар мен сандарды (тыныс белгілерін қоса) жояды [6]. Әрі қарай NLTK Python пакетін пайдаланып, мәтінді жеке сөздерге немесе таңбалауыштарға бөлу қажет, бұл мәтіндік деректерді сөздік таңбалауыш көмегімен оңай талдау және өңдеу. Сөзді жоюды тоқтату мәтіндік деректерді азайту және машиналық оқыту үлгілерінің дәлдігін арттыру үшін қолданылады. NLTK-дан WordNetLemmatizer сөздерді олардың негізгі мағынасын алу үшін олардың негізгі формасына қысқарту үшін пайдаланылады [7].

BM25 әдісі - бұл құжаттың белгілі бір сұранысқа қаншалықты сәйкес келетінін анықтау үшін қолданылатын ықтималдық моделі және рейтинг функциясы. BM25 «Үздік сәйкестік 25 (Best Match 25)» дегенді білдіреді. BM25 әр құжатқа сұранысқа қаншалықты сәйкес келетініне байланысты маңыздылық бағасын беру арқылы жұмыс істейді. Терминдердің жиілігі жоғары құжаттар қайта бағаланған кезде пайда болатын терминдердің жиілігінің шамадан тыс қанықтылығын болдырмау үшін алгоритм құжаттағы сұрау терминдерінің жиілігін де, құжаттағы сұрау терминдерінің кері жиілігін де ескереді. Corpus [8]. Ол көбінесе жетілдірілген әдістерді бағалау үшін анықтамалық алгоритм ретінде қолданылады. BM25FA - BM25 стандартының нұсқасы, ұзындықты қалыпқа келтіру, терминнің өзектілігімен қанықтылығы және құжаттың маңыздылығы әр түрлі болуы мүмкін көптеген өрістерден тұрады деген болжам бар-барлығы ескеріледі. BM25+ - BM25 стандартындағы кемшілікті түзететін тағы бір BM25 модификациясы, бұл сұрау терминіне сәйкес келетін ұзақ құжаттарды кейде әділетсіз түрде сұрау терминін мүлдем қамтымайтын қысқа құжаттарға бірдей сәйкес деп бағалайды. Бұл бір қосымша Delta тегін параметрін енгізеді.

Семантикалық іздеуде мәтіндік құжаттың мағынасын білдіру үшін Mean of Word vectors (MWV) тәсілі қарапайым және тиімді әдісті ұсынады. Бұл құжатты құжаттағы сөздердің векторлық орташалануы ретінде ұсынуды қамтиды. Сөз ендірудің алдын-ала дайындалған моделін қолдана отырып, құжаттағы әр сөз бастапқыда мәтіннің үлкен корпусы контекстінде әр сөздің семантикалық мағынасын білдіретін көп өлшемді вектор түрінде ұсынылады. Құжаттың бірыңғай векторлық көрінісін жасау үшін құжаттағы сөз векторлары орташаланады. Жеке сөздердің мәндеріне сүйене отырып, бұл вектор құжаттың жалпы мәнін көрсетеді.Word2vecalgorithm алгоритмі нейрондық желі моделін қолдана отырып, мәтіннің үлкен массивіндегі сөздердің корреляциясын зерттейді. Бұл лингвистикалық дискурста сөздердің контекстін қалпына келтіруге үйретілген екі қабатты, таяз нейрондық желі. Мәтіннің үлкен массивін кіріс ретінде қолдана отырып, Word2vec векторлық кеңістікті жасайды, әдетте бірнеше жүз өлшеммен және корпустағы әрбір жеке сөзге кеңістіктегі сәйкес векторды тағайындайды [9]. Word2vec Continuous Bag-of-Words (CBOW) моделінің

архитектурасын немесе Continuous Skip-Gram моделінің архитектурасын қолдана отырып, таратылған сөз көріністерін жасай алады. Ғаламдық векторлар немесе GloVe-бұл сөздерді ұсынудың үлестірілген моделі. Бұл сөздердің векторлық көріністерін құрудың бақыланбайтын әдісі және олардың арасындағы қашықтық олардың семантикалық жақындығымен байланысты болатын ыңғайлы кеңістікте сөздерді реттейді. Корпустағы сөздердің пайда болуының жаһандық статистикасына негізделген GloVe тренингінен алынған көріністер сөздердің векторлық кеңістігінің сызықтық ішкі құрылымдарын көрсетеді. FastText — Facebook Research компаниясы сөздерді ұсыну мен мәтінді санаттауды жылдам үйрену үшін әзірлеген кітапхана. Бақыланатын (классификациялар) және бақыланбайтын (енгізілетін) сөздер мен сөз тіркестерінің екеуі де FastText арқылы қолдау көрсетеді. Ол 157 түрлі тілге үйретілген үлгілерді ұсынады (сурет 1).

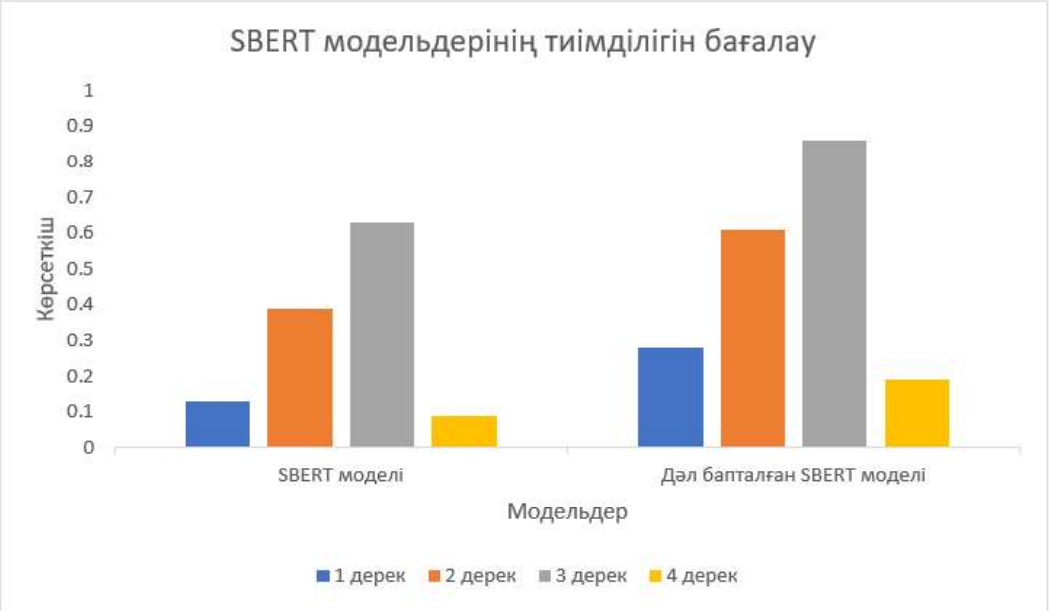



Сурет 1. MWV тәсілінің тиімділігін бағалау

Әмбебап сөйлемді кодтаушы (USE - The Universal Sentence Encoder), Google жасаған алдын ала дайындалған үлгі фразаның семантикалық мағынасын тұрақты ұзындық векторына кодтауға арналған. Сөйлем сөйлемнің тұрақты ұзындықты векторлық көрінісін беретін USE жүйесіне беріледі. Бұл векторды сөйлемдер арасындағы мағыналық ұқсастықты анықтау немесе құжаттар жинағынан ең ұқсас құжатты табу үшін пайдалануға болады. Алдын ала дайындалған Universal Sentence Encoder екі нұсқасы бар: біреуі Transformerencoder арқылы оқытылады және екіншісі Deep Averaging Network (DAN) арқылы оқытылады. Дәлдік және өңдеу ресурстарына деген қажеттілік екеуінің арасындағы келіссөздер болып табылады. «Трансформер» кодері бар нұсқа дәлірек болғанымен, ол көбірек есептеуді қажет етеді. DAN кодтауын пайдаланатын кодтың дәлдігі азырақ және есептеу құны жоғары [10].

Sentence-BERT (SBERT) стандартты алдын ала дайындалған BERT желісінің модификациясы болып табылады. Бұл сөйлемдердің семантикалық мағынасын тұрақты ұзындықты векторларға кодтайтын алдын ала дайындалған модель. Ол «Трансформер» архитектурасына негізделген және бақылаусыз оқыту әдістерінің әртүрлі ауқымын пайдалана отырып, мәтіннің үлкен корпусында оқытылады. Ол сиам және триплет желілерін пайдалана отырып, әрбір сөйлем үшін сөйлемді ендіруді жасайды, содан кейін оларды косинус ұқсастығы арқылы салыстыруға болады. Бұл сөйлемдердің үлкен санын семантикалық іздеудің орындылығына мүмкіндік береді (тек бірнеше секунд жаттығу уақытын қажет етеді). Пайдаланылатын сиамдық нейрондық желілер бір-біріне ұқсас екі немесе одан да көп ішкі желілерді қамтитын бірегей желілер болып табылады, екі модель

бірдей параметрлерді/салмақтарды ортақ пайдаланады және екі қосалқы модель де өз параметрлерін бірдей түрде жаңартады. SBERT-ті дәл реттеу үшін оқыту деректер әрбір жаттығу үлгісін сөйлем жұптарын білдіретін жолдар тізімі ретінде және олардың семантикалық ұқсастығын көрсететін белгімен сақтайтын InputExampleclass көмегімен дайындалады (сурет 2). Содан кейін стандартты PyTorch DataLoaderis осы жаттығу деректерін орау үшін пайдаланылады, ол деректерді араластыруға және белгілі бір топтамаларды жасауға мүмкіндік береді. өлшемі. Сөйлемдердің ұқсастығын тану үшін желіні дәл баптау үшін CosineSimilarityLoss қолданылады. Әрбір фразалық жұп үшін бұл жоғалту A және B сөйлемдерін SBERT желісі арқылы беру арқылы жасалған u және v кірістірулерінің ортақ ұқсастығын анықтайды. SBERT моделі бір дәуірге дайындалған және жаттығу қадамдарыдың алғашқы 10%-ы үшін қызады, деректер жүктеушісі және кіріс ретінде берілген жоғалту функциясы объектісі бар кортеждерден тұратын мақсаттарының тізімі.




Сурет 2. SBERT модельдерінің тиімділігін бағалау

*Нәтижелер және талқылау*

Бұл жұмыста BM-25 алгоритмін, сөз векторларының орташа мәнін, әмбебап сөйлемді кодтаушы үлгісін және семантикалық іздеу тапсырмасына арналған деректер жиынындағы сөйлем-BERT қарастырылады. Бұл зерттеу әртүрлі модельдердің тиімділігін бағалау үшін ақпаратты табу үшін офлайн бағалау көрсеткіштерін пайдаланады. Дербес көрсеткіштердің екі түрі бар: тапсырысты ескере отырып және тапсырысты есепке алмай. Ол жай ғана шынайы сәйкес нәтиже болжамды нәтижелерге енгізілгенін анықтайды; бұл шынайы сәйкес нәтиже болжамды нәтижелердің Топ-5 - 5 бірінші немесе бесінші орынға ие болғанына қарамастан бірдей нәтиже береді. Тәртіпті ескеретін көрсеткіштерді пайдаланған кезде болжамды Топ-5 нәтиженің бірінші позициясындағы шынайы мәнді нәтижеге бесінші позициядағы шынайы мәнді нәтижеге қарағанда үлкен балл беріледі. Бұл зерттеу үлгінің өнімділігін ретті ескермейтін көрсеткіштерді де, ретті ескеретін көрсеткіштерді де, атап айтқанда орташа өзара дәрежені пайдалана отырып салыстырады.

Семантикалық іздеу мәселесі үшін қарастырылған бірінші тәсіл – BM25, ол ықтималдық моделі болып табылады және семантикалық іздеу тұрғысынан TF-IDF сияқты басқа алгоритмдерден асып түседі. Құжаттың сұранысқа қаншалықты сәйкес келетінін анықтау

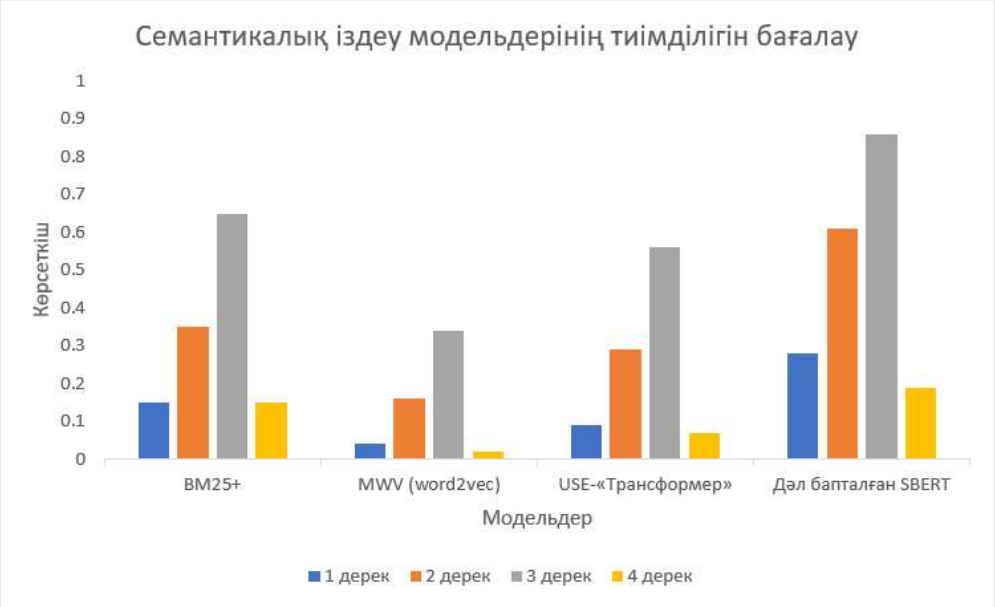
кезінде термин құжат ішінде пайда болатын контекст те, терминді корпустағы барлық құжаттарда қолданудың жалпы жиілігі де ескеріледі. Бұл сонымен қатар құжаттардың сұранысқа сәйкестігінің белгісіздігін түсіндіруі мүмкін. Құжаттардың сапасы мен өзектілігі әр түрлі болуы мүмкін үлкен және әр түрлі құжаттар жинағымен жұмыс істегенде, бұл өте пайдалы. BM-25 алгоритмінің нұсқалары арасында ең жоғары өнімділік моделі - BM-25 Plus. Векторлауға негізделген тәсілдер қарапайым және тиімді әдіс болумен қатар, олар бейімделгіш және адам тілінің болжамсыздығымен жақсы жұмыс істейді. Олар сондай-ақ мәтіннің семантикалық мағынасын жақсырақ түсіруге және адам тілінің әртүрлілігімен күресуге мүмкіндік беретін қарапайым және тиімді әдіс.

Қарастырылған екінші тәсіл - белгілі бір кемшіліктері болса да, қарапайым және тиімді сөз векторларын (MWV) пайдалану. Мысалы, ол мәтіндегі сөздер арасындағы байланысты немесе олардың ретін ескермейді. MWV модельдерін қолдану кейде олардың сөздік құрамына кірмейтін терминдермен жұмыс істей алмауымен шектеледі. Дегенмен, ең жақсы жұмыс істейтін нақты модель нақты тапсырмаға және пайдаланылатын деректерге байланысты.

Әрі қарай, Universal Sentence Encoder (USE) алдын-ала дайындалған модельдерін қарастырдық, олар мәтінмәндік ақпаратты кодтай алады, мысалы, сөз тәртібі мен сөйлемдегі сөздердің орналасуы, сөйлем құрылымы және сөздердің қатынасы, оларға нәзік мағынаны алуға мүмкіндік береді. Бұл оларға сөздік емес терминдерді немесе оқу деректерінде жоқ сөздерді өңдеуге және нәзік мағыналарды алуға мүмкіндік береді. Қолдану модельдері MWV-дің семантикалық іздеу тапсырмаларына деген көзқарасынан асып түсті, өйткені олар нәзік мағынаны қабылдауға, синтаксис пен сөздік емес сөздерді өңдеуге және көп тапсырмалы оқытудың пайдасын көруге қабілетті.

Sentence-BERT (SBERT) моделі «Трансформер» архитектурасына негізделген, ол сөздер мен сөйлемдер туралы контексттік ақпаратты түсіре алады, бұл семантикалық іздеу тапсырмалары үшін маңызды, өйткені мәтіннің мағынасын контексте түсінуді қажет етеді. SBERT нақты тапсырмаларды дәл баптауға арналғандықтан, ол нақты деректер сипаттамаларына бейімделе алады және сол тапсырмаларды орындау кезінде өнімділігін арттырады.

Алдын ала дайындалған SBERT және Дәл бапталған SBERT модельдері дәл баптау қабілеті, контекстік ақпаратты түсіру қабілеті және үлкенірек және әртүрлі деректер жиынтығында оқу қабілетінің арқасында семантикалық іздеу тапсырмаларында алдын ала дайындалған Universal Sentence Encoder (USE) үлгісіне қарағанда жақсы нәтиже көрсетті. Бұл зерттеуде екі модель ұсынылған: «Трансформер»-ге негізделген сөйлемді кодтау моделі және NLP басқа тапсырмаларына оқытуды тасымалдауды жеңілдету мақсатында сөйлемдерді ендірілген векторларға кодтауға арналған терең орташалау желісі (DAN) моделі. Ұсыныс деңгейінде алдын ала дайындалған ендірулерді қолданатын соңғы зерттеулер аударма тапсырмасының жоғары өнімділігін көрсетті. Зерттеу «Трансформер» (BERT) екі бағытты кодтаушы көріністері деп аталатын тілдік бейнелеу моделін әзірледі, ол таңбаланбаған мәтіннен терең екі бағытты көріністерді алдын ала үйрету үшін барлық деңгейлерде сол және оң контекстті бірлесіп анықтауға бағытталған. Ол табиғи тілді өңдеудің он бір операциясын орындау кезінде ең заманауи нәтижелерді қамтамасыз етеді. Белгілі бір тапсырмаға тән архитектураға елеулі өзгерістер енгізбестен, алдын ала дайындалған BERT моделі кең ауқымды тапсырмалар үшін озық үлгілерді әзірлеу үшін бір ғана қосымша шығыс деңгейімен одан әрі жетілдірілуі мүмкін. Bert (SBERT), косинустық ұқсастықты қолдана отырып салыстыруға болатын семантикалық мағыналы сөйлем тіркемелерін тудыратын алдын-ала дайындалған BERT желісінің модификациясы ұсынылған (сурет 3).




Сурет 3. Семантикалық іздеу модельдерінің тиімділігін бағалау

## Қолданылған әдебиеттер тізімі

1 Kuchuganov A.V. Semanticheskij analiz i poisk graficheskoj informacii: monografiya / Kuchuganov A.V. — Moskva : Aj Pi Ar Media, 2020. — 179 c. — ISBN 978-5-4497-0634-8, 31-78.

2 “Total data volume worldwide 2010-2025 / Statista.” https://www.statista.com/statistics/871513/worldwide-data-created/

3 Mäkelä E., “Survey of Semantic Search Research.” [Онлайн ресурс].

4 Wei W., Barnaghi P.M., and Bargiela A., “Search with Meanings: An Overview of Semantic Search Systems.” [Онлайн ресурс] - http://www.w3.org/TR/owl-guide/

5 Lee C.H., Noh H.R., Kim K.C., Design of Torque and Power Density Improvement According to the Rotor Shape of IPMSM. International Journal of Intelligent Systems and Applications in Engineering, 11(4s), 174–179. https://ijisae.org/index.php/IJISAE/article/view/2585.

6 Sudeepthi G., Anuradha G., M.B.-I.J. of Computer, and undefined 2012, “A survey on semantic web search engine,” Citeseer, 2012 [Онлайн ресурс].

7 Thakre B., Thakre R., Timande S., Sarangpure V., An Efficient Data Mining Based Automated Learning Model to Predict Heart Diseases. Machine Learning Applications in Engineering Education and Management – 2023, 1(2), 27–33.

8 Zhai C., Massung S., “Text Data Management and Analysis: A Practical Introduction to Information Retrieval and Text Mining June 2016 https://doi.org/10.1145/2915031.2915054dl.acm.org, [Онлайн ресурс].

9 Sherje D.N., Content Based Image Retrieval Based on Feature Extraction and Classification Using Deep Learning Techniques. Research Journal of Computer Systems and Engineering – 2021, 2(1), 16-22. https://technicaljournals.org/RJCSE/index.php/journal/article/view/14

10 Agrawal P., “Exploration of Proximity Heuristics in Length Normalization,” Jan. 2017, [Онлайн ресурс] - http://arxiv.org/abs/1701.01417.

## *Раздел 2. «Информационно-коммуникационные технологии»*

Садуакасов А.А., Мухаметжанова Б.О.

### Систематический и сравнительный анализ алгоритмов семантического поиска

В сегодняшнюю цифровую эпоху пользователям сложно найти нужную им информацию в Интернете из-за большого количества данных, которые легко доступны и постоянно генерируются ежедневно. Проблема связана с большим объемом данных, которые усложняют процесс поиска, а традиционные поисковые системы на основе ключевых слов могут быть неэффективны при обработке сложных запросов. В результате пользователи могут столкнуться с нерелевантными или неполными результатами поиска, что затрудняет поиск и получение необходимой им информации.Эту проблему можно успешно решить, используя семантический поиск, основанный на передовых методах машинного обучения и обработки естественного языка. Семантический поиск — это инновационный подход, который позволяет системам понимать значение и контекст пользовательских запросов. Этот метод позволяет не только учитывать отдельные ключевые слова, но и анализировать их смысловые связи и контекст, что в конечном итоге обеспечивает более точные и релевантные результаты поиска. Таким образом, семантический поиск становится ключевым инструментом, обеспечивающим пользователям более эффективный и приятный опыт поиска информации в онлайн-пространстве.

*Ключевые слова:* обработка естественного языка, семантический поиск, извлечение информации, поиск информации, большой объем данных, традиционные поисковые системы, векторы слов, производительность моделей.

A.A. Saduakasov, B.O. Mukhametzhanova

### Systematic and comparative analysis of semantic search algorithms

In today's digital age, users find it difficult to find the information they need online due to the large amount of data that is easily accessible and constantly generated on a daily basis. The problem comes from the large amount of data that complicates the search process, and traditional keyword-based search engines can be ineffective in handling complex queries. As a result, users may encounter irrelevant or incomplete search results, which makes it difficult for them to find and obtain the information they need. This problem can be successfully solved by using semantic search based on advanced techniques of machine learning and natural language processing. Semantic search is an innovative approach that allows systems to understand the meaning and context of user queries. This method allows not only to take into account individual keywords, but also to analyze their semantic relationships and context, which ultimately provides more accurate and relevant search results. Thus, semantic search becomes a key tool to provide users with a more efficient and satisfying experience in searching for information in the online space.

*Key words:* natural language processing, semantic search, information extraction, information retrieval, large data volume, traditional search engines, word vectors, model performance.

References

1 Kuchuganov A.V. Semanticheskij analiz i poisk graficheskoj informacii: monografiya / Kuchuganov A.V. — Moskva : Aj Pi Ar Media, 2020. — 179 c. — ISBN 978-5-4497-0634-8, 31-78.

2 “Total data volume worldwide 2010-2025 / Statista.” https://www.statista.com/statistics/871513/worldwide-data-created/

3 Mäkelä E., “Survey of Semantic Search Research.” [Онлайн ресурс].

4 Wei W., Barnaghi P.M., and Bargiela A., “Search with Meanings: An Overview of Semantic Search Systems.” [Онлайн ресурс] - http://www.w3.org/TR/owl-guide/

5 Lee C.H., Noh H.R., Kim K.C., Design of Torque and Power Density Improvement According to the Rotor Shape of IPMSM. International Journal of Intelligent Systems and Applications in Engineering, 11(4s), 174–179. https://ijisae.org/index.php/IJISAE/article/view/2585.

6 Sudeepthi G., Anuradha G., M.B.-I.J. of Computer, and undefined 2012, “A survey on semantic web search engine,” Citeseer, 2012 [Онлайн ресурс].

7 Thakre B., Thakre R., Timande S., Sarangpure V., An Efficient Data Mining Based Automated Learning Model to Predict Heart Diseases. Machine Learning Applications in Engineering Education and Management – 2023, 1(2), 27–33.

8 Zhai C., Massung S., “Text Data Management and Analysis: A Practical Introduction to Information Retrieval and Text Mining June 2016 https://doi.org/10.1145/2915031.2915054dl.acm.org, [Онлайн ресурс].

9 Sherje D.N., Content Based Image Retrieval Based on Feature Extraction and Classification Using Deep Learning Techniques. Research Journal of Computer Systems and Engineering – 2021, 2(1), 16-22. https://technicaljournals.org/RJCSE/index.php/journal/article/view/14

10Agrawal P., “Exploration of Proximity Heuristics in Length Normalization,” Jan. 2017, [Онлайн ресурс] - http://arxiv.org/abs/1701.01417.