10-дәріс

ҚАЗАҚСТАННЫҢ АШЫҚ УНИВЕРСИТЕТІ

ӘДЕБИЕТ ТЕОРИЯСЫ: АНТОЛОГИЯ, IV

Сандық гуманитарлық ғылымдар: зерттеу тәжірибесінің теоретизациясы

Мақсаты: Тед Андервудтың «Сандық гуманитарлық ғылымдар: зерттеу тәжірибесінің теоретизациясы» эссесімен және заманауи сандық технологияларды әдебиеттанушылық талдауда қолданудың мүмкіндіктерімен танысу.

Тірек сөздер: сандық технологиялар, мәтін, іздеу, талдау.

«Сандық зерттеулер – әдебиеттануда сандық мәтіндердің көмегімен бақылаулар мен болжамдарды дәлелдеуге мүмкіндік береді», – деп пайымдайды Тед Андервуд. Мәліметтерге интеллектуалды талдау жасай отырып, ғалымдар әртүрлі мәтіндегі белгілі бір сөздер немесе сөз тіркесін қалай қолдануға болатыны, оларды қолдану жиілігін анықтауға, әртүрлі мәтіндерді салыстыруға және ортақ белгілерін табуға мүмкіндік алды. Бұл үшін қағаз бен қарындаш қана керек. Өзінің эссесінде Тед Андервуд материалдарды іздеу және өңдеу тәсілі қалай жұмыс жасайтынын көрсетуге тырысты.

Кішігірім революция

Гуманитария сандық зерттеулер тәсілі туралы мәселені қолға ала бастады. Оған себеп – мұндай зерттеулердің бүгінде артта қалуы деп біледі Т. Андервуд. Мәліметтердің электронды қорын құру соңғы екі онжылдықта гуманитарлы ғылымдарда қарқынды қолға алынып келеді. Біз бұл үдерісті – «іздеу» деп атадық. Бірақ «іздеу» – күрделі технология үшін тым қарапайым атау. Интеллектуалды талдау аясындағы жаңа секілді көрінетін көптеген іс-әрекеттер біздің пәндерімізде білінбей орын тепті. Себебі гуманитарияда 1990 жылдардан бастап толықмәтіндік іздеу жұмыстары қолданыла бастады. Мәліметтерді интеллектуалды талдау – жаңа технология секілді танылып жүргенмен, Т. Андервуд оны біраз уақыттар бұрын басталған философиялық дискурс деп түсіндіреді. Бұл тәсіл алгоритмдік зерттеулерді жүзеге асыруға гуманитарияға көмек беріп келеді.

Біріншіден, іздеу ғылымда дәлелдеу рөлін атқарып жүр дегенді қалай түсінеміз? Мұнда «іздеу» сөзі әртүрлі технологиялар арасындағы шегараны жойып отыр. Библиографиялық іздеу – іргелі жұмыс болып табылады. Толықмәтінді іздеу де сол сияқты: біз тақырыпты ендіріп, іздеу жұмысын оп-оңай жүргізе аламыз. Бірақ екеуінің технологиясы мен тәсілдері әртүрлі.

Тәжірибеде толықмәтіндік іздеу – құжаттарды жинақтауға арналған логикалық мағыналық операция болып табылады. Андервуд «қызару» (ұялғаннан) сөзі XIX ғасырдағы адамгершілік сананың көрсеткіші болған деп топшылады. Сөйтіп бастапқы дереккөздерден «қызару» және «сана» сөздері қолданылған өлеңдер мен мәтіндерді іздей бастайды. Егер жоқ болса, іздеуді жалғастыра береді. Бірақ «қызару» сөзімен бірге «сана» емес, «ұялу» сөзін іздесе, нәтиже шығар ма еді?

Мұнда іздеу дегеніміз – материалды жылдам өңдеу ғана емес, эксперимент секілді мәнге ие тәсіл. Әрине, эксперимент әзірше нақты тәсіл болып отырған жоқ. Т. Андервуд таңдап алған іздеу терминдерін сөз немесе символдың әдеби мағынасы туралы болжам жасап, сол болжамын бірнеше сәтті талпыныстар нәтижесінде дәлелдейді. Ал шын мәнінде, зерттеуші мәліметтердің көп бөлігін қарастыру үшін алгоритм пайдаланып, іздеу үдерісі барысында материалды беру тәсілін қалыптастырып, ғалым дереккөздердің репрезентативтілі туралы интуитивті болжамдар жасаған.

Толық мәтінді іздеудің ішкі математикасы библиографиялық іздеуге қарағанда, мәліметтерді интеллектуалды талдау тәсілімен жақындығы бар. Егер Т. Андервудтың келтірген мысалына сүйенсек (ол «Моби Дик» атауы бойынша іздеу жүргізген), нәтижелерді оңай түсіндіруге болады. Бірақ толықмәтінді іздеуде көп жағдайда сәйкестіктер жиі кездеседі. Оның орнына алгоритм мәліметтерді белгілі бір релеванттылық деңгейіне сай сұрыптап шығуы керек.

Жалпы алғанда, толық мәтіндік іздеу – карталық каталогқа ұқсайтын іздеу құралы болып табылмайды. Толықмәтіндік іздеу – гуманитария бірнеше онжылдықтар бойы гипотезаларды тексеру және құжаттарды сұрыптау мақсатында қолданып келген алгоритмдердің үлкен тобының атауы. Толықмәтіндік іздеудің қарапайым формалары 1970 жылдардан бастап қолжетімді болып келеді. 1990 жылдарға дейін тарихи дереккөздердің CD-ROM мәліметтер қоры ешкімге танылмаған-ды. Бүгіннің өзінде бұл технология тарихи пәндерде қолданыл-

майды, себебі тарихшылар бұрын-соңды жарияланбаған мәтіндер мен дереккөздерге ізденіс жүргізуді қалайды. Алайда соңғы зерттеулерге сүйенсек, гуманитарияда іздеу жүйелері белгілі бір деңгейде қолданыла бастаған. Кез келген адам іздеу жұмыстарын Гуглдан бастайтын болған. Іздеу тәсілдерінің ғылыми нәтижелерін бағалау қиын. Себебі ғалымдар өздері ашқан жаңалықтарды жариялаған мақалаларында не еңбектерінде толық ашып беруден қашып, құпия ұстауды әдетке айналдырған.

Жаңа технологияның айқын нәтижесі – 90 жылдардағы әдебиеттанушылар өздерінің алғашқы дереккөздердің кең диапозонында жазатынын байқағаны. Т. Андервуд ғылыми мәселелерді құрастыру аясы да өзгергенін, себебі олар енді толықмәтінді іздеу тәсілінің мүмкіндіктерін пайдалана бастағанын жазады. Айталық, 1990-жылдары кейбір тақырыптар Т. Андервуд үшін қиынға соққан: Конгресс кітапханасында «британ романтикалық дәуіріндегі энергия» мәселесіне арналған пәндік каталог болмады. Толықмәтіндік іздеу бұл тақырыпты зерттеуге жеңіл етіп тастады. Тақырыпты вербалды белгілермен байланыстырып жіберсек, оннан астам цитаталар табылып, тақырып «дискурс» екенін анықтала салады екен.

Т. Андервуд бір кездері қиналғанына ұялатынын да айтады. Зерттеу ережесінің оңайға өзгергені сонша, ешқандай зерттеуші енді қиындық көрмейтін болады. Белгілі бір тақырыптың маңыздылығын анықтау үшін қанша дереккөзді қарау керек? Жиырма ма? Ал толықмәтіндік іздеуде оның саны екі не үш есеге жоғарылауы мүмкін. Миллиондаған сілтемелерден толықмәтінді іздеу тақырыпқа нақты сай келетін жиырма сілтемені бөліп ала алады.

«Жерасты тастары»

Толықмәтіндік іздеу зерттеуші жұмыс жасап отырған кез келген тезисті дәлелдей алатынын рас. Бірақ мәселе мұнымен бітпейді. Сұрыптау дереккөздері сіз жұмыс жасамаған балама тезистерді де сұрыптай тастайтыны бар. Іздеу – мәліметтерді интеллектуалды талдау формасы, бірақ, ол сіз бұрыннан білетін мәліметтерді қайталап көрсетуі де, сіз күткен нәтижелерді сұрыптап беруі де мүмкін. Гуманитария бұл мәселені 1990 жылдарға дейін талқылауға ниеттенген де жоқ, себебі іздеу жүйелері – ірі сандық коллекцияларға қолжетімділікті қамтамасыз ететін жалғыз тәсіл еді. Алайда соңғы жылдары зерттеу тәжірибелері әртараптанып, іздеудің герменевтикалық шектеулері айқын көріне бастады.

Компьютерлік ғылымдар саласының ғалымдары қажетті ақпаратты мәліметтердің үлкен қорынан алуға мамандана бастады (1). Олар интерпретацияны философиялық қатаң ескеруге негізделген іздеу жолдарын құрды. Т. Андервуд «философиялық қатаң ескеру» сөзінің өзі – нағыз жаңалық болуы мүмкін екенін жазады. Гуманитария информатиканы – философиялық емес, құралдық дискурс деп қарауға үйренген. Мәліметтердің интеллектуалды талдаудың негізгі тілі – байесов статистикасы – интерпретация туралы тұжырымдау тәсілі, бізге мәтіндер қорын ұстанымды тәсілдер арқылы зерттеуге мүмкіндік беретін жол (2, 3). Байесов статистикасы – герменевтикалық шиыршық секілді. Гуманитария бұл статистикамен жақсы таныс, себебі біз әрбір мәселені алдын ала болжамдай отырып, талдаймыз. Жаңа дәлелдермен бетпе-бет келген тұста, біз болжамдарды бір-бірімен сәйкестендіріп, оларды белгілі бір деңгейде қайта анықтай бастаймыз. Мәселе бір мәтін туралы болса бұл герменевтикалық стиль интуицияға негізделер еді. Интеллектуалды талдау бір ғана оқырман қарастыра алмайтын мәтіндердің көп қорымен қалай жұмыс жасай алатынын түсіндіреді. Мәліметтерді салыстырудың барлық стратегиялары бізге белгілі бір деңгейде шаблонды құруға негіз болады. Бірақ кейбір стратегиялардың болжамдар теріске шығаратын дәлелдерді табуға мүмкіндігі бар. Мәселен, әдебиеттанушылардың көпшілігі тақырыптарды анықтау үшін тірек сөздерді іздей бастайды (айталық, қызару/ұялу немесе жұмыс/энергия). Бірақ бұл жерде сөздердің сәйкес келуі олардың мағыналары арасындағы байланысты анықтайды деп қателеседі. Бұл тілшілер «орналастыру гипотезасы» деп атаған мағына үлгісімен байланысты. Әрине, аталған үлгінің маңызы зор деп айта алмаймыз, бірақ информатикада ғана емес, әдеби зерттеулерде де эвристика ретінде пайдалануда көмегі бар. Орналастыруға негізделген алгоритм белгілі бір кезеңде қолданылған терминмен байланысты тілді анықтауға қабілетті болады. Айталық, «қызару» сөзімен ассоциацияланатын сөз XVIII және XIX ғасырдағы 4820 жинақтың мәтіндерінде кездескен

екен. Бірақ тірек сөздерді пайдаланып, іздеу жүргізу барлық мақсаттарды орындайды дегенді білдірмейді. Егер сіз не іздеп отырғаныңызды жақсы білсеңіз, іздеу жүйесі – таптырмас құрал бола алады. Алайда тарихи ғылымда біз не іздеп отырғанымызды білмей, қателесетін кездер болатыны дәлелденген.

Шын мәнінде, деп пайымдайды Т. Андервуд, – біздің іздеп отырған тақырыбымыз тек бір ғана сөзбен байланысты болмауы мүмкін. Мәселен, қызару сөзімен. Бәлкім, біздің тақырыпты ашу үшін басқа терминді қолдану керек болған шығар?! Бәлкім, белгілі бір тақырыпқа қатысты әлеуметтік құбылыстар бірнеше терминдердің көмегімен ашылар ма еді?! Сондықтан, біз баспа жазбаны бір-бірімен салыстырып, алгоритмге терминдер кластерін құруға мүмкіндік беруіміз керек. Бұл процесс зерттеуші іздеуді ойламаған дискурсивті шаблондарды анықтауға көмек береді.

Тақырыптарды үлгілеу сөзді бірнеше тақырыпқа қатысты етіп, олардың арасындағы ассоциация заңдылығын құруы мүмкін. «Күлкі» лексемасы (және одан туындаған сөздер) XVIII және XIX ғасырлар поэзиясында әртүрлі тақырыпта қолданылғанын көреміз. Поэзияның 13 798 томын Т. Андервуд XVIII және XIX ғасырлардағы 469 000 томнан HathiTrust электронды кітапханасынан іріктеп алған. Осындай көлемді жинақты поэзияны өңдеу үшін ғалым жанрлық салыстыру тәсілін пайдаланады (6).

Әрине, әр тақырыпты бөлек-бөлек қарастыруға болар еді, яғни «күлкі» сөзі бір тақырыптан екінші тақырыпта уақытқа қарай ауысып отыратынын көрсетуге болады. Алгоритм «күлкі» сөзі жиі қолданылған 117 тақырыпты анықтап берген. Әрбір тақырып бойынша белсенді қолданылған сөздер тізімі жасалып, ақындар «күлкі» сөзін пайдаланған контекстерге талдау жасалды. «Күлкінің» XVIII ғасыр поэзияларының көпшілігіндегі сатиралық функциясы мен сентименталды немесе аматорлық сипаты («тәтті», «таза», «көз» ұғымында) арасында қарама-қайшылық байқалды.

«Күлу», «күлкі» секілді сөздер 13 789 томда тақырыптар бойынша жіктелген. 120 тақырыптан Т. Андервуд үш тобын бөліп алды. Мұнда «күлкі» сөзі белсенді қолданылады.

Т. Андервуд бұл өзгерістер арасындағы себеп-салдардық байланыс туралы айтып тұрған жоқ. Ғалымның иллюстрациясы тақырыптық үлгілеу бақылау туралы тұжырым жасауға түрткі болған. Бірақ бақылау нәтижесі терең нәтижелерге сеп болуы мүмкін. Мәтью Джокерс тақырыптық үлгілеуді XIX ғасыр романдарын құру үшін пайдаланады; Роберт К. Нельсон бұл терминді Азаматтық соғыс жылдарындағы газеттердің тақырыптық акценттерін көрсету үшін пайдаланған. Эндрю Голдстоун мен Т. Андервуд бұл техниканы XX ғасырдағы әдебиеттанудың әртүрлі сыни терминдерін анықтау мақсатында қолданған.

Тақырыптық үлгілеу туралы сөз еткеннен гөрі Т. Андервуд алгоритмдік зерттеу туралы әңгіменің маңыздылығына назар аударады. Тақырыптық үлгілеу пікірталасқа негіз болған, болып қала бермек.

Зерттеушілер алгоритмдерді жабық әрі құпия «қара жәшік» ретінде қарай алмайды. Егер біз алгоритмді өз зерттеулерімізде қолданатын болсақ, онда олардың қалай жұмыс жасайтынын анықтап алу керекпіз. Тақырыптық үлгілеу – патенттелмеген жаңалық болып саналады. Тақырыптық үлгілеу алгоритмі – жалпыға қолжетімді, сондықтан гуманитария оларды өздерінің мақсатына қарай қолданып, өзгерте алуға толық құқылы.

Алгоритмнің интерпретациялық шектеулерін түсіну үшін оның математикалық негізін түсіну керек. Айталық, тақырыптық үлгілеудің кең таралған формасында тақырыптардың саны – үлгілеу процесіне ендірілген болжамдардың бірі. Алгоритм кез келген дискурс немесе оның шегараларының бірлігі туралы нақты жауап бере алмайды. Алгоритм ассоциациялардың заңдылықтарын анықтауда қолжетімді әрі маңызды рөл атқара алады.

Алгоритмдерді зерттеу үшін пайдалану бірқатар философиялық мәселерді туындатты. Гуманитария сандық тәсілдерді қолданғысы келеді. Алгоритмді құрал ретінде пайдалану сиқыр көрсету секілді жаңалық. «Қазіргі тілдер ассоциациясының басылымдарында» Алан Лю үлгілеушілер (үлгілеу тәсілін жиі пайдаланушылар) «нәсіл» кестесін сәйкестендіру мақсатына қол жеткізе алды ма, жоқ па деген сұрақ туындайтынын айтады.

Егер шын мәнінде сондай болса, философиялық тығырыққа тірелеріміз анық. Журналдарда біздің тығырыққа тірелгенімізді дәлелдейтін технофилді рапсодияларды табуға болады. Алайда бұл рапсодиялар статистикалық үлгілер туралы жеткілікті деңгейде ақпараттанбаған. Тақырыптық

үлгілеу (немесе мәліметтерді интеллектуалды өңдеудің басқа да алгоритмдері) «болжамсыз» жұмыс жасайды дегенді білдірмейді. Үлгі – есептеу техникасы теориясының мамандары мен адамдары құрған абстракция. Байесов үлгісі қазіргі пәндерде жиі қолданылып келеді.

Тақырыптың үлгіні құжаттардың жиынтығына ендіргісі келген зерттеуші іздеп тапқысы келген тақырыптарды және нені тапқысы келетінін білуі тиіс. Үлгілеу үдерісінде компьютер жоқтан идея жасамайды. Егер өнер саласындағы терминдерді пайдаланатын болса, компьютер бізге үлгіні дәлелге «жуықтата алады». Зерттеудің бұл тәсілі тірек сөздерді іздеуге қарағанда, әлдеқайда ашық болуы мүмкін. Бірақ интерпретация үдерісі адамның болжамдарын қалыптастыра алады.

Т. Андервуд гуманитария мәліметтердің көп қорын талдай алатынын жазады. Мәселе мынада, – деп пайымдайды ол, – біз іздеу алгоритмін проекциялау тәсілі ретінде қолданып, историзмге қарсы әрекет ете бастадық. Информатиканың статистикалық тілі біздің пәндік дәстүрімізге жат болуы мүмкін, – деп ойлайды Т. Андервуд, – шындық мынада: гуманитария зерттеу тәсілдерін жасауы үшін осы тілді түсінуі керек. Сөйтіп біз мәліметтердің үлкен қорымен жұмыс жасай алуына мүмкіндік аламыз.

Бұл – гумунитария үшін пәнаралық сұхбаттың жаңа түрі, соның нәтижесінде біз көптеген пайдалы дүниелерді біле аламыз. Солай ете отырып, өз үлесімізді қосамыз. Т. Андервуд нақты пәндерде герменевтикалық теорияның маңызды үлгісі болатынын жазады. Үлгілеудің қиындығы мынада – тарихи өзгерістер гуманитарлы ғылымдарға түсінікті бола бермейді. Баспа құжаттарын белгілі бір уақыт шеңберінде үлгілеуге тырысқан ғылымдар үздіксіздік туралы болжамдар жасайды. Сөйтіп, ғылыми әдіснама мен гуманитарлы теорияның өнімді алмасуына мүмкіндік туындайды.

Тақырыпты бекітуге арналған сұрақтар

1. Көркем мәтінді дәстүрлі және сандық өңдеудің арасындағы айырмашылық неде?
2. Әдебиеттану ғылымында сандық технологияның жаңа мүмкіндіктері қандай?
3. Мәтіндердің көп бөлігін сандық өңдеу тәсілінің кемшіліктері мен артықшылықтары қандай?

Әдебиеттер

1. Coenen Frans Data Mining: Past, Present, and Future// Knowledge Engineering Review. – 2011. – № 26.
2. Bovens Luc, Hartmann Stephan. Bayesian Epistemology. – Oxford, 2003.
3. Kruschke K. John. Doing Bayesian Data Analysis. – Burlington, MA, 2011)
4. Sahlgren Magnus. The Distributional Hypothesis//Rivista di Linguistica. – № 20.
5. Turney D. Peter, Pantel Patrick. From Frequency to Meaning: Vector Space Models of Semantics// Journal of Artificial Intelligence Research. – 2010. – № 37.
6. Underwood Ted. Mapping Mutable Genres in Structurally Complex Volumes: International Conference on Big Data-2013//arXiv.org (Cornell University Library, http://arxiv.org/abs/1309.3323).
7. Jockers L. Matthew. Macroanalysis: Digital Methods and Literary History. – Urbana, 2013.
8. Nelson K. Robert. Mining the Dispatch, Digital Scholarship Lab// University of Richmond http://dsl.richmond.edu/dispatch/.
9. Goldstone Andrew, Underwood Ted. The Quiet Transformations of Literary Studies: What Thirteen Thousand Scholars Could Tell Us//New Literary History (forthcoming) https://www.ideals.illinois.edu/handle/2142/49323.
10. Schmidt M. Benjamin. Words Alone: Dismantling Topic Models in the Humanities//JDH: Journal of Digital Humanities. – 2012. http://journalofdigitalhumanities.org/2-1/words-alone-by-benjamin-m-schmidt/
11. Wang Xuerui, McCallum Andrew. Topics over Time: A Non-Markov Continuous-Time Model of Topical Trends//Proceedings of the Twelfth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. – New York, 2006.