Дәріс 2

Data Mining Ұғымы. Деректерді когнитивті талдау. Big Data ақпарат көздерін шолу

Data Mining – бұл сұраныстарға сәйкес ақпаратты қолданушыларға ұсыну, ұйымдастыру, сақтау, толықтыру және қолдау үшін арналған автоматтандырылған жүйе.

Деректерді талдау – кең ұғым. Бүгінгі таңда оның ондаған анықтамалары бар. Ең жалпы мағынада деректерді талдау – бұл көптеген параметрлері бар көп өлшемді жүйені есептен шығарумен баланысты зерттеу. Деректерді талдау барысында осы деректер арқылы сипатталатын қандай да бір көрсеткіштердің пайда болу тарихын анықтау үшін зерттеуші белгілі бір іс-әрекеттер орындайды. Әдетте, деректерді талдау үшін әр түрлі математикалық әдістер пайдаланылады.

Деректерді талдауды ақпаратты жинап болған соң, оны тек ақпаратты өңдеу ретінде ғана қарастыруға болмайды. Деректерді талдау - бұл, ең алдымен, гипотезаларды тексеру құралы және зерттеушінің міндеттердін шешу.

Адамның мүмкіндігі шектеулі танымдық қабілеттері мен Ғаламның шексіздігі арасындағы белгілі қарама – қайшылықтары бізді модельдер мен модельдеуді қолдануға итермелейді, осылайша бізді қызықтыратын нысандарды, құбылыстар мен жүйелерді зерттеу оңайға түседі.

Data Mining технологиясының мәні мен мақсатын былайша тұжырымдауға болады: бұл – айқын емес, объективті және тәжірибе жүзінде пайдалы заңдылықтары бар үлкен көлемді деректерді іздеуге арналған технология.

Айқын емес заңдылықтар – бұл ақпаратты өңдеудің стандартты әдістерімен немесе сараптау жолымен табуға болмайтын заңдылықтар.

Объективті заңдылықтың астында, әрқашан субъективті болып табылатын экспертті пікірден ерекшеленетін, толығымен шындыққа сәйкес келетін заңдылықтарды түсіну қажет.

Бұл деректерді талдау тұжырымдамасы келесіні болжайды:

- Деректер нақты емес, толық емес, қарама-қайшы, әртекті, жанама, және соның өзінде үлкен көлемді болуы мүмкін; сондықтан нақты қосымшалардағы деректер түсінігі елеулі зияткерлік күш-жігерді талап етеді;
- Деректерді талдау алгоритмдерінің өздері " ақыл-ой элементтеріне" ие болуы мүмкін, атап айтқанда, прецеденттер бойынша оқу қабілеті, яғни жеке бақылаулар негізінде жалпы қорытындылар жасау; мұндай алгоритмдерді құру сондай-ақ елеулі зияткерлік күш-жігерді талап етеді;
- Шикі деректерді ақпаратқа өңдеу процестері, ал ақпараттар білімге қолмен орындалуы мүмкін бола алмайды, және автоматтандыруды талап етеді.

Data Mining технологиясының негізіне деректердегі көпаспектілі өзара қарым-қатынастың фрагменттерін көрсететін үлгілер (паттерндер) концепциясы салынған. Бұл үлгілер жинақы және адамға түсінікті түрде болатын деректердегі сынамаларды алуға тән заңдылықтарды білдіреді.

Іздеу шаблондары сынамаларды алудың құрылымы туралы априорлы жорамалдардың шектеусіздігімен және талданатын көрсеткіштер мәндерін бөлу түріндегі әдістермен жүргізіледі.

Data Mining технологиясының маңызды ерекшелігі ретінде ізделінетін шаблондардың стандартты еместігі және айқын еместігі болып табылады.

Басқаша айтқанда, Data Mining құралдарының OLAP құралдар мен деректерді статистикалы өңдеу құрал-жабдықтарынан келесідей ерекшеленеді: қолданушылармен өзара тәуелділікте алдын ала болжанатын тексеру орнына, олар қолда бар деректер негізінде мұндай тәуелділікті өз бетінше табуға және олардың сипаты туралы гипотеза құруға қабілетті.

Data Mining әдістерімен анықталатын заңдылықтардың стандартты бес типтерін бөліп көрсетеді:

- Қауымдастық (association) – оқиғалардың бір-бірімен байланысының жоғарғы ықтималдығы. Қауымдастықтың мысалы ретінде дүкендерде жиі бірге сатып алынатын тауарларды айтуға болады;
- Реттілігі (sequence) – оқиға уақытымен байланыста болатын тізбектің жоғарғы ықтималдығы. Реттіліктің мысалы ретінде бір тауарды сатып алғаннан кейін, белгілі бір кезең ішінде басқа тауарды сатып алу ықтималдығы жоғары болатын жағдай бола алады;
- Жіктеу (classification) – кандай да бір оқиға немесе нысан тиесілі болатын топты сипаттайтын белгілері болады;
- Кластерлеу (clustering) – жіктеумен ұқсас заңдылық және одан айырмашылығы – топтардың өздері берілмейді, олар деректерді өңдеу процессі кезінде автоматты түрде анықталады;
- Болжау (forecasting) – сол немесе өзге деректердің мінез-құлық динамикасындағы үлгілердің бар болуы. Болжаудың сипатты мысалы – қандай да бір тауар немесе қызметке деген сұраныстың маусымдық өзгеруі.

Data Mining мақсаттары. Заманауи Data Mining компьютерлік термині «ақпарат алу» немесе «деректерді өндіру» деп аударылады. Data Mining сөзімен қатар Knowledge Discovery («білім табу») және Data Warehouse («деректер қоймасы») терминдері жиі кездеседі. Data Mining-тің ажырамас бөлігі болып табылатын, жоғарыда көрсетілген терминдердің пайда болуы деректерді сақтау және өңдеу әдістері мен құралдарының дамуындағы жаңа бағдарымен байланысты. Сонымен, Data Mining мақсаты үлкен көлемді (өте үлкен) деректердің жасырын ережелері мен заңдылықтарын анықтаудан тұрады. Себебі, адамның ақыл-ойы өзімен өзі орасан зор алқаптағы әртүрлі ақпаратты қабылдау үшін бейімделмеген. Орта есеппен адам, кейбір жеке тұлғаларды есептемегенде, тіпті шағын таңдаулар ішіндегі екі-үш өзара байланысты

қабылдауға қабілетті емес. Бірақ сонымен қатар ұзақ уақыт бойы деректерді талдаудың негізгі құралы рөліне үміткер болған дәстүрлі статистика да нақты өмірден алынған міндеттерді шешу кезінде жиі тоқтап қалады. Ол жиі жалған шамалар болып табылатын таңдаудың орташа сипаттамасын басқарады (клиенттің орта төлем қабілеттілігі, мұнда тәуекел немесе шығын функциясына байланысты сізге клиенттің ниеті мен жағдайын болжауды үйрену қажет; сигналдың орташа қарқындылығы, мұнда сізді сигналдың ең жоғарғы шегі мен алғышарттарының сипаттамасы қызықтырады).

Сондықтан математикалық статистика әдістері негізінен алдын-ала тұжырымдалған гипотезаны тексеру үшін пайдалы болады, ал гипотезаны анықтау кейде жеткілікті күрделі және көп еңбекті қажет ететін тапсырма болып табылады.

Data Mining – бұл жалғыз емес, білімді табудың әртүрлі әдісінің үлкен сандар жиынтығы. Әдісті таңдау жиі қолда бар деректердің түріне және басқа қандай ақпарат алуға тырысатыңызға байланысты. Мысалы кейбір әдістер: қауымдастық, классификация, кластерлеу, уақытша қатар талдауы және болжау, нейронды желілер және т.б.

Анықтамада берілген білімнің қасиетін толығырақ қарастырайық.

Білім бұрын сонды белгілі болмаған, жаңа болуы керек. Қолданушыға бұрыннан белгілі болған білімді ашуға жұмсалған күш-жігер ақталмайды. Сондықтан да құндылықты тек қана жаңа, бұрын белгісіз болған білім береді.

Білім тривиальды емес болуы керек. Талдау нәтижелері, жасырын білім дегенді құрайтын, деректердегі айқын емес, күтпеген заңдылықтарды көрсетуі тиіс. Неғұрлым қарапайым тәсілдермен алынған нәтижелер (мысалы, көзбен көрумен) Data Mining қуатты әдістерін тартуды ақтамайды.

Білім тәжірибе жүзінде пайдалы болуы керек. Табылған білім қолданылуы керек, соның ішінде сенімділігі жеткілікті жоғары дәрежеде болатын жаңа деректерде де қолданылуы тиіс. Оның пайдалылығының мәні – бұл білімдер оларды қолдану кезінде белгілі бір пайда әкелуі болып табылады.

Білім адамның түсінуіне қол жетімді болуы керек. Табылған заңдылықтар логикалық түсінікті болуы тиіс, олай болмаған жағдайда олар кездейсоқ болады деген ықтималдық бар. Сонымен қатар табылған білім адам үшін түсінікті түрде берілуі тиіс.

Data Mining-те алынған білімді ұсыну үшін модельдер қолданылады. Модель түрлері оларды құратын әдістерге тәуелді. Ең көп таралған болып табылады: ережелер, ағаштар шешімдері, кластерлер және математикалық функциялар [1].

Data Mining қолданылу аясы ештеңемен шектелмеген – қандай да бір деректер болатын жердің барлығында Data Mining керек. Көптеген кәсіпорындар тәжірибесі көрсеткендей, Data Mining қолдану арқылы 1000% қайтарым алуға болады. Мысалы, бастапқы шығындар 350-ден 750 мың долларға 10-70 есе асып түскен экономикалық әсер туралы хабар белгілі. Небәрі

4 ай ішінде ақталып шыққан 20 млн. долларлық жоба туралы мәліметтер келтіріледі. Басқа мысал – Ұлыбритания универсам желілеріне Data Mining құралын енгізу есебінен жылдық үнемдеу 700 мың доллар болды. Data Mining жетекшілер мен талдаушылар үшін олардың күнделікті қызметінде үлкен құндылықты ұсынады. Іскер адамдар Data Mining әдістерінің көмегімен олар бәсекелестік күресте елеулі артықшылықтарды алатындарын түсінді.