

Дәріс 6. Оптималды кодтар құру. Деректерді қысу үшін Хаффман әдісін қолдану.

Дәріс жоспары.

1. Оптималды кодтар құру.
2. Деректерді қысу.
3. Деректерді қысу үшін Хаффман әдісін қолдану.

Оптималды кодтар құру.

Оптималды кодтар құру - ақпарат теориясының ең маңызды міндеттерінің бірі. Бұл бағыттың мақсаты - ақпаратты мүмкіндігінше аз битпен, бірақ қателіксіз және түсінікті түрде жеткізуді қамтамасыз ету. Оптималды код – бұл орташа кодтық сөз ұзындығы минималды болатын код, яғни хабарды қысқаша әрі тиімді ұсыну тәсілі. Мұндай кодтарды құру барысында әрбір хабар элементінің ықтималдығы, энтропия ұғымы және кодтың құрылымдық қасиеттері ескеріледі.

Оптималды кодтаудың теориялық негізін Клод Шеннон қалаған. Оның ақпарат энтропиясы туралы теоремасына сәйкес, кез келген ақпарат көзін кодтау кезінде орташа код ұзындығы энтропиядан кем болмайды. Энтропия $H(X)$ - хабардағы белгісіздік немесе орташа ақпарат мөлшерін сипаттайтын шама. Сондықтан оптималды кодтың орташа ұзындығы $L_{\min} \approx H(X)$ шамасына барынша жақын болуы тиіс. Осылайша, энтропия кодтау тиімділігінің теориялық шегі болып саналады.

Оптималды кодтарды құру үшін бірнеше маңызды талаптар бар. Біріншіден, код бірімді декодталатын болуы керек, яғни әрбір кодтық тізбек тек бір хабарға сәйкес келуі тиіс. Екіншіден, код префикстік болуы керек — яғни, ешбір кодтық сөз басқа кодтық сөздің басы болмауы қажет. Бұл талаптар хабарды қате түсіндіруді болдырмайды және кодты автоматты түрде талдау мүмкіндігін береді.

Оптималды кодтарды құруда ықтималдық негізіндегі тәсілдер қолданылады. Әрбір хабар элементінің пайда болу ықтималдығы неғұрлым жоғары болса, соған сәйкес код қысқарақ болуы керек. Бұл принцип Хаффман кодтау әдісі мен Шеннон–Фано әдісінде жүзеге асырылады. Осылайша, хабардың жиі кездесетін бөліктері аз битпен, ал сирек кездесетіндері ұзын кодпен бейнеленеді. Нәтижесінде жалпы код ұзындығы азаяды, яғни код тиімділігі артады.

Мысалы, хабар көзіндегі төрт символдың ықтималдықтары $A:0.4, B:0.3, C:0.2, D:0.1$ болса, оптималды кодтау нәтижесінде AAA мен BBB символдары қысқа кодтар (мысалы, 0 және 10), ал CCC мен DDD ұзын кодтар (110 және 111) алады. Бұл тәсіл жалпы хабарды кодтауға жұмсалатын бит санын айтарлықтай азайтады.

Оптималды кодтардың тағы бір маңызды қасиеті — олардың экономикалық тиімділігі. Егер кодтық сөздердің ұзындығы символдардың ықтималдығына пропорционалды болса, онда арнаның өткізу қабілеті барынша тиімді пайдаланылады. Бұл принцип телекоммуникацияда,

мәліметтерді сақтау жүйелерінде және бейне-аудио компрессияда ерекше маңызға ие.

Кодтау тиімділігін бағалау үшін код тиімділігі (η) және артықтылық (R) ұғымдары қолданылады. Код тиімділігі энтропия мен орташа код ұзындығының қатынасымен анықталады.

Оптималды кодтарды құру тек деректерді қысуға ғана емес, сонымен қатар ақпарат беру жүйелерінің өнімділігін арттыруға мүмкіндік береді. Бұл кодтар байланыс арнасындағы жүктемені азайтады, сақтау көлемін үнемдейді және хабарды беру жылдамдығын арттырады. Сонымен қатар, дұрыс құрастырылған кодтар қате түзету жүйелерімен бірігіп, ақпараттың сенімді берілуін қамтамасыз етеді.

Қазіргі таңда оптималды кодтау принциптері көптеген технологияларда қолданылады. Мысалы, Хаффман кодтау ZIP, JPEG, MP3 форматтарында; Арифметикалық кодтау бейне және аудио компрессияда; Lempel–Ziv алгоритмдері файлдық жүйелер мен интернет деректерін қысу хаттамаларында пайдаланылады. Бұл әдістердің барлығының негізінде оптималды код құру идеясы жатыр.

Қорытындылай келе, оптималды кодтар құру — ақпарат теориясының өзегі болып табылады. Мұндай кодтар энтропиямен тығыз байланысты және ақпаратты жіберу, сақтау мен өңдеу процестерін барынша тиімді етуге бағытталған. Оптималды кодтаудың дамуы қазіргі заманғы цифрлық технологиялардың — желілердің, мультимедиа жүйелерінің және бұлттық есептеулердің — негізін құрайды.

Деректерді қысу.

Деректерді қысу — ақпарат теориясының және информатиканың маңызды саласы, оның басты мақсаты — деректерді сақтау мен жіберуге қажетті көлемді азайту. Қысу кезінде бастапқы ақпараттың мәні сақталады, бірақ ол ықшам түрде бейнеленеді. Бұл процесс әсіресе жады шектеулі құрылғыларда, желілер арқылы мәлімет алмасуда және мультимедиа файлдарын өңдеуде үлкен рөл атқарады. Деректерді қысу әдістері ақпараттың құрылымына, қайталанатын үлгілеріне және ықтималдық қасиеттеріне сүйенеді.

Қысу әдістері негізінен екі түрге бөлінеді: жоғалтусыз қысу (lossless compression) және жоғалтумен қысу (lossy compression). Жоғалтусыз қысуда бастапқы деректер толық қалпына келтіріледі. Мұндай тәсілдер мәтіндік файлдар, бағдарламалық кодтар, немесе цифрлық құжаттар үшін қолданылады. Ал жоғалтумен қысу кезінде ақпараттың кейбір бөлігі жойылады, бірақ ол адамның қабылдауында байқалмайды. Бұл әдіс бейне, сурет және дыбыс сияқты мультимедиялық деректерде кеңінен қолданылады, себебі олардың дәлдігіне аздаған өзгеріс енгізу сапаны айтарлықтай бұзбайды.

Жоғалтусыз қысуға мысал ретінде Хаффман кодтау, Шеннон–Фано кодтау, Lempel–Ziv (LZ77, LZ78) және Run-Length Encoding (RLE) әдістерін келтіруге болады. Бұл алгоритмдер деректер ішіндегі қайталанатын үлгілерді, жиі кездесетін символдарды және ықтималдық байланыстарды анықтайды. Мысалы, Lempel–Ziv алгоритмінде бұрын кездескен деректер бөліктері сөздік

түрінде сақталып, кейін соларға сілтеме арқылы қайта қолданылады. Осылайша, бірдей тізбектердің қайталануының алдын алып, көлем айтарлықтай азаяды.

Жоғалтумен қысу әдістерінің ең танымал мысалдары — JPEG (суреттерге), MP3 (дыбыстарға) және MPEG (бейнеге) форматтары. Бұл тәсілдер адамның көру және есту жүйесінің ерекшеліктерін пайдаланады: мысалы, көз кейбір ұсақ түстер айырмашылығын немесе құлақ белгілі бір жиіліктегі дыбыстарды байқамайды. Осы қасиеттерді ескере отырып, алгоритм маңызды емес ақпаратты алып тастайды, нәтижесінде файл көлемі бірнеше есе азаяды.

Деректерді қысудың тиімділігі энтропия ұғымымен тікелей байланысты. Энтропия — ақпараттағы белгісіздік өлшемі, яғни бір хабарды бейнелеу үшін қажетті минималды орташа бит саны. Егер деректердегі энтропия төмен болса (мысалы, қайталанулар көп болса), онда оларды қысу жеңіл және тиімді болады. Керісінше, энтропия жоғары болса, яғни ақпарат кездейсоқ болса, қысу мүмкіндігі шектеулі болады.

Қысу әдістерінің басты артықшылығы — жад пен өткізу қабілеттілігін үнемдеу. Мысалы, қысылған файл аз орын алады және желі арқылы тезірек беріледі. Бұл әсіресе IoT құрылғыларында, мобильді жүйелерде және бұлттық есептеу орталарында өте маңызды. Сонымен қатар, қысу энергия тұтынуды да азайтады, себебі аз деректерді өңдеу мен жіберу аз энергия қажет етеді.

Дегенмен, қысу кезінде кейбір шектеулер де бар. Жоғалтумен қысуда ақпарат сапасы төмендеуі мүмкін, ал жоғалтусыз қысуда қысу дәрежесі шектеулі. Сонымен қатар, қысу мен ашу (декодтау) процестері есептеу ресурстарын қажет етеді, сондықтан жүйенің өнімділігі мен ресурстары да ескерілуі тиіс.

Қысу технологиялары тек файлдар үшін емес, сонымен қатар желілік хаттамаларда, мәліметтер базасында, және бейнелерді ағындық жіберу жүйелерінде қолданылады. Мысалы, бейне конференциялар кезінде бейне мен дыбыс нақты уақытта қысылып, арна арқылы жіберіледі. Бұл деректердің тасымалын жылдам және тұрақты етеді.

Қазіргі уақытта деректерді қысу бағытында жасанды интеллект пен машиналық оқыту әдістері де қолданылуда. Нейрондық желілер бейнелер мен дыбыстардың құрылымдық ерекшеліктерін үйреніп, оларды тиімдірек кодтау жолдарын таба алады. Мұндай тәсілдер классикалық алгоритмдерге қарағанда анағұрлым жоғары қысу деңгейіне жетуге мүмкіндік береді.

Қорытындылай келе, деректерді қысу — ақпараттық технологиялардың ажырамас бөлігі. Ол деректерді сақтау, тасымалдау және өңдеу процестерін оңтайландырады, жүйелердің тиімділігін арттырады және ресурстарды үнемдеуге мүмкіндік береді. Болашақта қысу әдістері жасанды интеллект пен ықтималдық модельдеріне сүйене отырып, одан да жоғары тиімділікке жетеді деп күтілуде.

Деректерді қысу үшін Хаффман әдісін қолдану

Деректерді қысу үшін Хаффман әдісін қолдану — ақпарат теориясындағы ең кең тараған және тиімді тәсілдердің бірі. Бұл әдістің негізгі

идеясы — хабардағы әрбір символдың пайда болу ықтималдығына қарай оған қысқа немесе ұзын код тағайындау. Яғни, жиі кездесетін символдарға қысқа, ал сирек кездесетін символдарға ұзын код беріледі. Нәтижесінде орташа код ұзындығы азайып, деректерді сақтау мен жіберуге қажетті көлем қысқарады. Хаффман әдісі жоғалтусыз қысу (lossless compression) санатына жатады, яғни бастапқы ақпарат толық қалпына келтіріледі.

Хаффман кодтау әдісін 1952 жылы америкалық ғалым Дэвид Хаффман ұсынған. Ол бұл әдісті Шеннон мен Фаноның ықтималдыққа негізделген кодтау тәсілдерін жетілдіру мақсатында жасаған. Хаффман алгоритмі оңтайлы кодтарды автоматты түрде құруға мүмкіндік береді және дәлелденген түрде ең қысқа орташа ұзындықты код береді. Сондықтан ол қазіргі заманғы көптеген файлдық және мультимедиялық қысу жүйелерінің негізін құрайды.

Хаффман кодтау процесі бірнеше кезеңнен тұрады. Бірінші кезеңде бастапқы хабардағы барлық символдар анықталып, олардың пайда болу жиілігі немесе ықтималдығы есептеледі. Екінші кезеңде осы символдардан бинарлық ағаш құрылады: ең аз ықтималдыққа ие екі символ біріктіріліп, олардың қосынды ықтималдығы бар жаңа түйін пайда болады. Бұл процесс барлық символдар бір түбірге біріктірілгенге дейін қайталанады.

Ағаш толық құрылғаннан кейін үшінші кезең — код тағайындау кезеңі басталады. Ағаштың әрбір тармағына 0 және 1 мәндері жазылады. Түбірден әрбір жапыраққа дейінгі жол — сол символдың Хаффман коды болып саналады. Нәтижесінде жиі кездесетін символдар ағаштың түбіне жақын орналасады (қысқа код алады), ал сирек кездесетіндері тереңірек орналасады (ұзын код алады).

Мысалы, егер хабар “ABBCDA” түрінде болса және символдардың жиілігі: A – 2, B – 2, C – 1, D – 1 болса, онда ең аз ықтималдықты C мен D біріктіріледі. Содан кейін олардың қосындысы B және A түйіндерімен біріктіріледі. Нәтижесінде, мысалы, $A \rightarrow 0$, $B \rightarrow 10$, $C \rightarrow 110$, $D \rightarrow 111$ кодтары алынуы мүмкін. Осылайша, бастапқы 6 символды көрсету үшін 12 бит қажет болса, қысылған түрінде ол 10 битпен ғана бейнеленеді.

Хаффман әдісінің басты артықшылығы — оның оптималдығы мен бірімәнді декодталуы. Хаффман кодтары әрқашан префикстік код болып табылады, яғни ешбір код басқа кодтың басы бола алмайды. Бұл қасиет декодтау процесін жеңілдетеді және хабарды қате түсіндірудің алдын алады.

Декодтау процесі де өте қарапайым. Кодталған биттер ағаштың түбірінен бастап оқылады: әрбір 0 немесе 1 таңбасы сәйкес тармаққа апарады, ал жапыраққа жеткен кезде тиісті символ анықталып, нәтижеге қосылады. Осылайша, хабар бірімәнді түрде қалпына келтіріледі және деректердің ешбір бөлігі жоғалмайды.

Хаффман кодтау әдісі көптеген қолданбалы салаларда қолданылады. Мысалы, ZIP, GZIP сияқты файл қысу форматтарында, JPEG сурет кодтауында, MP3 дыбыс қысуында және PDF құжаттарында. Бұл алгоритмнің қарапайымдылығы мен тиімділігі оны ақпараттық жүйелер мен телекоммуникация саласында стандартқа айналдырды.

Алайда, Хаффман әдісінің кейбір шектеулері де бар. Егер хабарда барлық символдардың ықтималдығы бірдей болса, онда қысу тиімділігі төмендейді. Сондай-ақ, шағын көлемді немесе кездейсоқ деректер үшін Хаффман кодтау айтарлықтай үнем бермеуі мүмкін. Мұндай жағдайларда арифметикалық кодтау сияқты жетілдірілген әдістер қолданылады.

Қорытындылай келе, Хаффман әдісі — деректерді жоғалтусыз қысудың ең тиімді және математикалық тұрғыдан дәлелденген тәсілдерінің бірі. Ол ықтималдыққа негізделген құрылымы арқасында хабардың артық ақпаратын азайтып, ақпаратты сақтау мен тасымалдауды оңтайландырады. Хаффман кодтау — қазіргі заманғы деректерді өңдеу мен компрессия жүйелерінің іргетасын қалаған классикалық әрі әмбебап әдіс болып табылады.

Пайдаланылған әдебиеттер тізімі

1. Шеннон, К. Э. *Математикалық коммуникация теориясы*. — М.: Наука, 1963.
2. Хаффман, Д. А. *A Method for the Construction of Minimum-Redundancy Codes*. — Proceedings of the IRE, 1952.
3. Фано, Р. М. *The Transmission of Information: A Statistical Theory of Communications*. — Cambridge, MIT Press, 1961.
4. Коберн, Д., Трембле, Г. *Теория кодирования и ее приложения*. — М.: Мир, 1982.
5. Маю Венбо. *Прикладная криптография. Протоколы, алгоритмы и исходные тексты на языке С*. — М.: Триумф, 2004.