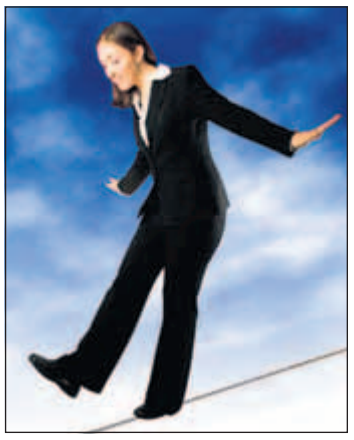
IDEER weekendavisen
ANNO 1749
Nr. 17 25. - 29. APRIL 2008 REDAKTØR: TINE EIBY

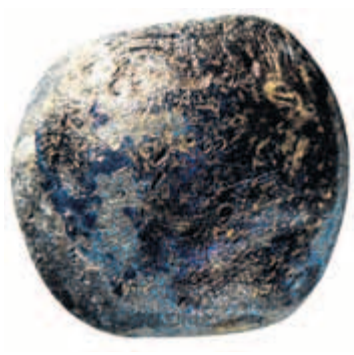


Arbejdsliv

Vi balancerer mellem almagt og afmagt
Side 6

Kvægpest

Ingen tør erklære kampen for vundet
Side 9



Perlefund

Gravet frem af Edens Have
Side 8

Tvivl!

Den dydige krigsmagt tager nemt fejl
Side 4

Massive data. Forestil Dem 37.000 kongresbiblioteker gange 17 millioner bøger fulde af bogstaver. Vi har langt flere data, end vores computere kan klare. Danske forskere arbejder på en løsning:

Logaritmen af *n*

Jeg ensø-

Af RENÉ GUMMER

ENHVER, der har haft en computer mellem fingrene, ved, at det næsten er en naturlov, at den bryder sammen en gang imellem. Besynderligt nok sker det altid, mens man har mest travlt, eller har glemt at gemme det, man sidder og arbejder med. Pludselig fryser skærmen, og dér røg månedens regnskab eller det romanlange brev til onkel Bob i Amerika, og det eneste, man får som forklaring til at trøste sig med, er en uforståelig meddelelse fra Windows.

Hvis man arbejder på et netværk, er det endnu værre. Pludselig holder tastaturet op med at virke, skærmen fryser, man bliver smidt af netværket, og så stopper dagens arbejde, indtil en eller anden fra computerafdelingen finder ud af, at det er hende den langhårede fra bogholderiet, der har forsøgt at hive mere data ned på sin pc, end netværket kan håndtere og på den måde har overbelastet serveren.

Det er blandt andet den slags problemer, professor Lars Arge sidder og tumler med. Som chef for Danmarks Grundforskningsfonds *Center for Massive Data* på Aarhus Universitet forsøger han og hans medarbejdere at regne sig frem til en bedre løsning til at håndtere den stadigt stigende datamængde og -størrelse.

»Overordnet set laver vi algoritmer, der løser problemer på den mest effektive måde. De mest fundamentale problemer er søgning og sortering af data. Det lyder måske lidt langhåret, men i virkeligheden er det meget enkelt: Forestil dig for eksempel, at du skal finde en, der hedder Jensen i telefonbogen. Du kunne begynde helt forfra og bladre dig gennem a'erne, b'erne, c'erne og så videre, indtil du kommer til J, og endelig finder den Jensen, du har brug for. Metoden skal nok lede dig frem til Jensen, men det tager lang tid. Det kaldes en lineær søgning.«

Lars Arge og hans medarbejdere vil ændre den søgemetode helt grundlæggende.

»Vi forsøger at lave en algoritme, der kan gøre det mere intuitivt, ligesom de fleste af os ville gøre det i en rigtig telefonbog – hvis der da overhovedet er nogen, der kigger i en telefonbog nu om stunder. Man slår op sådan cirka i midten og finder ud af, om man er kommet for langt eller for kort i forhold til navnet Jensen. Lad os så sige, at midten er ved K. Så er du kommet for langt og ved, at Jensen i hvert fald ikke er at finde i den sidste halvdel af bogen. Allerede der har du halveret den lineære søgningstid ved simpelthen at udelukke halvdelen af alle navne i bogen.

Så koncentrerer man sig om den første halvdel og gør det samme igen: slår op i midten og konstaterer nu, at man er havnet ved E og altså ikke langt nok. Nu kan du så fokusere på den anden halvdel og har, på samme søgningstid som det ville tage at kigge tre navne igennem med den lineære metode, udelukket tre fjerdedele af telefonbogen. Og sådan fortsætter du med at udelukke halvdele, indtil Jensen er fundet. Det kaldes en binær søgning, og det går meget, meget hurtigere end at kigge hele telefonbogen igennem. Algoritmen er en slags opskrift på, hvordan man søger mest effektivt.«

– Hvad så hvis man ikke har tusind, men en million navne i telefonbogen?

»Sådan nogle som mig kan godt lide at sige, at der er *n* navne i telefonbogen i stedet for at bruge konkrete tal som tusind eller en million. Det tager altså *n* tid at hente et stykke data, hvis du bruger den lineære metode. Hvorimod hvis du søger binært, så tager det logaritmen af *n* tid – vi kalder det *log n*. Du deler *n* med to så mange gange, at du til sidst er nede på én. Hvilket lige præcis er logaritmen

FORTSÆTTES SIDE 2

LAY-OUT: BENTE BRUUN, KORREKTUR: FLEMMING GERTZ

FORTSAT FRA FORSIDEN

Logaritmen af *n*

af *n*. Med den lineære metode er det absolut ikke lige meget, om der er tusind eller en million navne, men det er stort set underordnet, hvor stor *n* er med den binære algoritme. *n* vokser meget hurtigere, end *log n* vokser.«

EFFEKTIV søgning er allerede et nøgleord. I 2002 opgjorde seniorforskere ved Berkeley Universitet i Californien, USA, den samlede mængde af tilgængelige data alene på internettet til at være omtrent fem exabytes. Det er et hårrejsende stort antal data. For at give en idé om, hvor meget det drejer sig om, beder forskerne om, at man forestiller sig de omkring 17 millioner bøger i Kongressens bibliotek, hvor hvert eneste bogstav repræsenterer et stykke data. For at komme op på fem exabytes, skal man bruge 37.000 af den slags biblioteker. Det var i 2002, og man anslår, at tallet bliver fordoblet cirka hvert andet år. Så der er god grund til at kunne søge effektivt og systematisk.

LARS Arges arbejde begynder helt nede i bunden af systemet.

»For at kunne designe en ordentlig algoritme, er man nødt til at opstille en matematisk model af en computer. Man er nødt til at have en model for, *hvad* en maskine er, og hvad den kan. Helt traditionelt siger man, at en computer består af en processor og uendelig hukommelse. Det er den matematiske definition af en computer, vi ofte designer vores software efter i dag. Men vi *har* ikke uendelig hukommelse – *ram* – på vores maskiner. Ram er begrænset til mellem en og fire gigabyte – otte, hvis det går rigtig højt.«

Det er langtfra nok, når man gerne vil arbejde med store mængder data.

»Problemet opstår, fordi den model af en computer er meget dårlig, når man skal håndtere store datamængder. Grunden er, at dine data er nødt til at ligge ude på harddisken og ikke inde i ram-hukommelsen, hvor der simpelthen ikke er plads. En harddisk er en mekanisk tingest med en magnetisk plade, der skal køre rundt, og en læse-skrive arm, der skal flytte sig. Lidt ligesom en gammeldags grammofon. Den er mekanisk, og det betyder per definition, at den er langsom, for læsearmen skal altså flytte sig rent fysisk.

Selve hukommelsen – ram – er meget hurtigere, faktisk en million gange hurtigere. Det vil i praksis sige, at det er en million gange hurtigere at hente et stykke data, der ligger i din ram i forhold til, hvis du skal finde samme data på harddisken. Derfor er det selvfølgelig lidt underligt, at vi stadig arbejder efter en model, der siger, at en maskine består af en processor og uendelig meget hukommelse, for det har vi ikke i realiteten. Og når man har data på harddisken, der er større, end maskinen kan håndtere, så bryder tingene sammen.«

Netop fordi harddiske er så langsomme, er de designet sådan, at de faktisk læser mere end det stykke data, man har bedt om.

»Det kunne jo være smart at hente nogle flere data, der hvor læsearmen er nu. Det kan gøres ret hurtigt sammenlignet med den tid, det har taget læsearmen at komme frem til rette sted. Det er faktisk det, der sker i praksis. Når du siger, at du gerne vil have et stykke data, så får du rent faktisk en kæmpe klump, bare for en sikkerheds skyld. Som regel alt det, der kan transporteres – helt op til 64 kilobyte data – kommer med over i din ram-hukommelse. Det kaldes *block-access*.«

Det hjælper bare ikke så meget, hvis man ikke har tænkt sig at bruge de data, der er blevet hentet med over som bonus.

»Det ville jo være smartere, hvis man kunne designe en algoritme, der – lige efter at den har givet dig det, du har bedt om – begynder at arbejde på noget af det, den har fået gratis med over.«

IDEEN er, at når man alligevel har siddet og ventet på sine data, kan man lige så godt nyde godt af den ekstra bonusklump, der har taget turen med over i ram-hukommelsen. På den facon bliver den gennemsnitlige ventetid kortere.

»Vores tilgang er, at bruge en matematisk model hvor man har en processor og en *begrænset* mængde hukommelse. Og ved siden af en harddisk, der er uendelig stor. Vores algoritmer går så ud på at kigge så få gange som muligt på harddisken og *loade* så få som muligt af de her blokke ind i ram-hukommelsen for at løse et problem. Det drejer sig 0om at læse eller skrive så få blokke som muligt. For hver gang man gør det, koster det tid.«

Arbejdet med at finde på gode algoritmer, er en langsommelig proces, hvor man prøver sig frem. Professor Arges arbejde stopper dog ikke, når han har designet en perfekt algoritme. Han skal også bevise, at han har gjort det.

»Typisk designer vi flere forskellige algoritmer, efterhånden som vi bliver klogere. Dem analyserer vi så, forsøger at beskrive dem matematisk for at finde ud af, hvor lang tid det tager dem at finde gennem telefonbogen – for nu at blive i det eksempel. Når vi så har designet en algoritme, der finder Jensen hurtigst muligt, kigger vi på, hvor mange gange der skal kigges i telefonbogen, før Jensen er fundet, for at være sikre på, at han ikke kan findes hurtigere.

Så vores arbejde er både at designe algoritmen og analysere, hvor hurtig den er. Det gør også, at vi kan sammenligne algoritmerne, og at vi har et mål for, hvor god en algoritme er. For hvis man sidder som programmør og laver en algoritme, så kan det godt være, at man kan overbevise sin chef om, at man har lavet noget rigtig godt, men så vil han sikkert spørge, om man ikke kan lave en, der er endnu bedre. Og så er det vigtigt at kunne bevise matematisk, at det ikke kan lade sig gøre; at man har lavet den bedste algoritme, der kan laves til den givne opgave.«

DEN teknologiske udvikling gør, at mængden af data er hastigt stigende. Det stiller anderledes og større krav til de computere, der skal bearbejde dem. Professor Arge fortæller:

»Jeg arbejder for eksempel meget med terrændata. For ikke så lang tid siden fik man billeder af jordoverfladen fra for eksempel satellitter, til at kortlægge kloden med højder og det hele. Den metode gav det, vi kalder 100-meter-data. Det vil sige, at for hver hundrede meter, kender man jordoverfladens højde. Nu om stunder bruger man laserscanner.

Lad os nu tage Danmark. Her har et firma laserscannet hele landet med et punkt per meter. Det giver rigtig meget data i forhold til den gamle metode, lad os sige en terrabyte. Når der er så meget data, er det ikke et spørgsmål om, at det går langsomt, når man regner på det; der er simpelthen ikke noget software, der er hurtigt nok til at regne på det. På grund af størrelsen er data nødt til at ligge på harddisken, og så tager det så lang tid, at det slet ikke kan betale sig at bruge de data. Årsagen er, at de programmer, der kan regne på terrændata, er lavet på et tidspunkt, hvor man brugte 100-meter data. Dengang kunne man nogenlunde sige, at man havde lige så meget ram-hukommelse, som man havde data, og derfor bekymrede man sig ikke om, at programmet skulle kunne arbejde med et medie, der var en million gange langsommere end ram-hukommelsen. Derfor kan man rent faktisk ikke bruge laserscanningerne til særlig meget.«

I dag er problemet altså ikke at finde data at analysere. Faktisk er der så meget tilgængeligt datamateriale, at meget af det slet ikke kan bruges til noget.

»Det er lidt omvendt af, hvad det var for bare få år siden. Det, du vil analysere, er ikke længere begrænset af mængden af data, men af, hvor store mængder data, din maskine kan håndtere, typisk summet op i, hvor meget der kan være i din ram-hukommelse.«

DET er ligegyldigt, om man sidder ved en super-computer eller med en gammel, skrammet laptop. Store datamængder skal ses i forhold til den konkrete maskine, man arbejder på, og Lars Arges algoritmer kan bruges af både store og små. Men hvorfor ikke bare fylde noget mere ram-hukommelse i maskinen?

»Fordi det med at købe mere ram er lidt ligesom at tisse i bukserne om vinteren. Det virker et stykke tid, men så er man tilbage til start. Hvis du for eksempel har to gigabyte ram, så kan du godt doble det op og løse et fire-gigabyte problem. Men det varer ikke særlig længe, så vil du gerne løse et otte-gigabyte problem, fordi datamængderne bliver ved med at vokse.«

Løsningen kan ifølge Arge kun findes i at ændre for eksempel måden at søge på.

»At købe mere ram-hukommelse eller en hurtigere maskine betyder bare, at du kan kigge på måske ti navne på samme tid, som du før brugte på at kigge på ét. Den lineære algoritme er stadig en fundamentalt dårlig måde at søge i telefonbogen på. Selv om vi ikke er kommet så langt endnu, så er det i teorien teknisk muligt at fylde en terrabyte ram i din maskine, men det bliver meget hurtigt ekstremt dyrt og meget uhandy. Der kommer hele tiden nye teknologier, der gør, at vi får mere og mere regnekraft, men min tese er, at man på den måde bare midlertidigt løser et større og større problem. Man skal stadig hele vejen gennem telefonbogen, hver gang man skal finde Jensen.«

FALSIFICERET

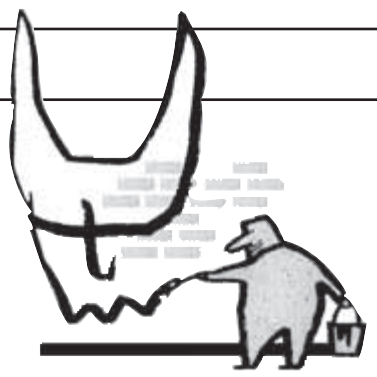
Tunge børn

DET var ikke for at kunne bære deres tunge babyer i armene, at fortidsmenneskene, hominiderne, rejste sig på to ben, hævder forskere fra University of Manchester. Det har ellers været en af flere anerkendte begrundelser. I modsætning til abeunger har menneskebørn ikke været udstyret til i samme grad at kunne klynge sig til deres ophav, ligesom menneskebørn har været længere tid om at lære at gå selv.

Men en ny undersøgelse af kvinders energiforbrug, når de bærer på en plump genstand på op til 10 kilo, fastslår, at denne oprejste bæreposition er den dårligst tænkelige anvendelse af energi, og at det derfor er utænkeligt, at evolutionen har favoriseret en sådan model.

eiby

University of Manchester, 23. april



1-0 til hjernen

SÅ er der nyt til den tilbagevendende diskussion om, hvorvidt mennesket har en fri vilje. For nu har forskere lavet en undersøgelse, som viser, at hjernen er markant foran, forstået på den måde, at den beslutter sig for handlinger op til syv sekunder før, man tager den bevidste beslutning.

Dermed tilbagevises forestillingen om, at vi har en fri vilje, konstaterer chefforskeren bag undersøgelsen, John-Dylan Haynes fra det Berlin-baserede Bernstein Center for Computational Neuroscience. Undersøgelsen gik ud på, at forskerne bad 14 forsøgspersoner om at trykke på to knapper, når de selv ville. Samtidig var MRI-scannere koblet til forsøgspersonernes hjerner, og man kunne der i det enkelte tilfælde konstatere, at den såkaldte planlægningshjerne – den præfrontale cortex – lyste op omkring syv sekunder før, fingeren rent faktisk trykkede på knappen.

Tidligere forsøg – ved blandt andre Nobelpristageren Benjamin Libet – har påvist en tidsforskydning mellem hjerneaktivitet og fysisk handling på 300 millisekunder.

Her mangler dog en meget væsentlig diskussion af, hvorvidt hjernens aktiviteter ikke også repræsenterer individets frie vilje.

eiby

New Scientist, 19. april

Arktisk dis

FORURENINGSPARTIKLER fra industriområder har en kedelig evne til at sprede sig med luftstrømmene til fjerne egne, og i det arktiske område kan de således være årsagen til en ret uappetitlig dis. Opdagelsen af fænomenet tilskrives J. Murray Mitchell, der var meteorolog i det amerikanske luftvåben, og som i 1957 beskrev sine hyppige observationer af disen. Senere målinger viste, at den indeholdt tungmetaller og andre stoffer, som typisk stammer fra afbrænding af tung olie.

En forskergruppe fra universitetet i Utah har nu til deres egen overraskelse fundet ud af, at mennesket var i stand til at forurene polarområdet næsten 100 år tidligere. Ved at gennempløje litteraturen er de stødt på adskillige beretninger fra polarekspeditioner, som beskriver tågen og et samtidigt gråt eller sort lag støv på isen.

Blandt kilderne er den svenske geolog Adolf Erik Nordenskiöld, som efter en rejse til Grønland berettede om sine iagttagelser i tidsskriftet *Science* i 1883. Allerede under en tidligere ekspedition i 1870 havde Nordenskiöld imidlertid observeret et gråt lag støv, som blev sort eller brunt, når det blev vådt, og som lå på indlandsisen i et lag på mellem 0,1 og 1 millimeters tykkelse. Han konstaterede også, at støvet indeholdt små mængder af jern, kobolt samt nikkel, og han formodede derfor, at det drejede sig om »kosmisk støv«, som var kommet hertil ude fra verdensrummet.

Desværre mener forskerne fra Utah nu, at Nordenskiölds kosmiske støv har en noget mere jordnær forklaring, nemlig forurening fra datidens industriskorstene. På grund af den ineffektive teknologi i 1800-tallet er det endda muligt, at partikelforureningen i Arktis har været større dengang end i dag, men det kan ændre sig igen med den kraftige vækst i de nye og mindre miljøbevidste industrilande som Kina.

jopp

Bulletin of the American Meteorological Society, marts 2008

LAY-OUT: BENTE BRUUN, KORREKTUR: FLEMMING GERTZ