

Neurale informatieopslag inspiratiebron voor het DBMS

Biologische Databases

Rick van Rein

In ons dagelijks leven werken we met strakke, stabiele en uiterst voorspelbare databases. Toch is het leerzaam om ook eens een uitstapje te maken naar een wat minder vast omliggende database, namelijk de inhoud van onze hersenpan.

Over de werking van onze hersenen is veel bekend, met uitzondering van het allerhoogste niveau, waarop de psychologie nog zoekende is. Maar we weten hoe hersencellen werken, we weten hoe ze leren en we weten vrij goed wat denken inhoudt. Dat goedbekende deel wordt in dit artikel afgezet tegen een alledaags DBMS. Want een dergelijke parallel kan soms uiterst stimulerend werken in de ontwerpsituatie.

In de hersenwerking is ruwweg een paar lagen te onderscheiden. Op de onderste laag, zeg maar de fysieke laag, bevinden zich de chemische processen die informatie overdragen en opslaan. Een standaardstelling van dergelijke processen vormt een neuron, waarvan ook alternatieve implementaties bestaan in de vorm van computergesimuleerde bits. Neuronen leren informatie die aangeeft wanneer ingespeeld moet worden op een situatie in de omgeving. Die omgeving bestaat uit invoer van andere neuronen; in samenwerking van een cluster neuronen zijn hier stabiele patronen in te vormen. De opeenvolging van dergelijke patronen heten associaties, en die kunnen we zien als de gedachtegang: denken is een opeenvolging van associaties.

Neuronen zijn lerende cellen

Een neuron of zenuwcel is de meest elementaire eenheid in het brein. Het heeft meerdere ingangen (die dendrieten heten) en een enkele uitgang (die axon heet). De axon werkt als een draad, die vrij lang kan zijn en die aankomt bij de dendrieten van andere cellen. Natuurlijk sluiten sommige dendrieten aan op ingangssignalen zoals de trilhaartjes (cilia) in onze oren; aan het andere uiteinde sturen sommige axonen actuatoren zoals spieren aan. Maar alles wat daartussen gebeurt valt onder de hersenwerking. Neuronen reageren op het gecombineerde ingangssignaal, in plaats van op een enkele ingang. Wel is het zo dat er verschil zit in het enthousiasme voor de verschillende ingangen, zodat de ene ingang wat sneller tot een respons leidt dan een andere. Dit zijn voorkeuren die zijn aangeleerd.

Als een neuron op zijn uitgang 'vuurt', en dat leidt tot een positieve bekrachtiging, dan leert de neuron dat door het verhogen

van het enthousiasme voor de ingangen die aanzet gaven tot het vuren. Dit betekent in de praktijk dat de verbinding tussen de ingang en zijn herkomst wordt verdikt, wat de elektrochemische weerstand verlaagt.

Op dit niveau is aan neuronen het een en ander te leren. Ze kunnen bijvoorbeeld eenvoudige, contextvrije conversies maken. Een typische praktijkoefening is bijvoorbeeld om een netwerkje te leren om een hexadecimale code (in 4 bits) om te zetten in de afzonderlijk aan te sturen signalen van een 7-segments display. Geen bijster spannende toepassing, maar wel iets dat betrouwbaar aan een neurale netwerk geleerd kan worden. Merk op dat het gaat om een digitaal probleem; doordat neuronen vuren of stil blijven is het domein waarover neuronen gaan altijd digitaal. Wanneer we neuronen als schakelingen gaan zien zoals die ook in de elektronica voorkomen, dan verliezen we echter een belangrijk aspect uit het oog, en dat is timing. Zoals we later zullen laten zien is dat van belang voor een goed begrip. Neuronen sparen als het ware impulsen aan de ingang op, tot het moment dat ze voldoende hebben om te vuren. Eenmaal gevuld heeft een neuron tijd nodig om weer tot zichzelf te komen, of eigenlijk: om lading op te bouwen.

Holistische database

Wie een neuron probeert te zien als een opslagplek voor een miniem beetje informatie, zal aannemen dat een query naar die informatie geïmplementeerd kan worden met de juiste prikkels op de ingangen van de database van al die neuronen. Nu is dat wat erg simpel, want niet alle neuronen kunnen op alle prikkels reageren, er is maar een beperkt aantal ingangen per neuron. Er zit vaak wat routing tussen om al die prikkels te transporteren van bijvoorbeeld onze ogen naar onze mond. Maar als geheel bezien is uit een netwerk van neuronen wel informatie te halen door de juiste prikkels aan te bieden. Als het bijvoorbeeld een ja/nee-vraag betreft dan wordt hetgeen waarover geoordeeld moet worden aangeboden op vaste ingangen, waarna het antwoord uit te lezen is op een neuron dat ook gebruikt is als uitgang tijdens het trainen.

Bizar aan deze vorm van data-opslag is dat we niet kunnen aanwijzen waar de informatie zich bevindt. In een database zouden we een sector op een schijf, en daarbinnen een offset, kunnen bepalen waar we de data vinden. De opslag is holistisch, ofwel de opslag ligt verspreid over cellen die elk meerdere informatie-

brokjes opslaan. Er is dus sprake van redundantie, maar dan wel over opslagplekken die niet louter en alleen dat ene stukje informatie opslaan. Maar de cellen samen weten wel al die afzonderlijke stukjes informatie terug te vormen als ze op de juiste wijze geprikkeld worden.

Denken is een opeenvolging van associaties

Zonder deze 'holistische' manier van dataopslag zou het brein ook niet kunnen functioneren; als een bit aan informatie in een enkele cel zou worden opgeslagen dan zou dat beetje verloren gaan zodra die cel zich een beetje aanpast in een leerproces. De enige uitzondering is eigenlijk de uitgangscel waar zo-even een ja/nee-signaal van werd opgepikt; dit is een cel die is aangewezen door het leerproces, en die verbindingen aangaat met voorgaande cellen om tot de juiste conclusie te komen. Doordat die cel niet wordt gebruikt voor andere doeleinden, is dat wel een cel die aanwijsbaar een stuk informatie representeert.

In context geplaatst

Een neuron krijgt *inputs* van diverse plaatsen; als daar maar positieve bekrachtiging voor bestaat dan wil een neuron met plezier allerhande verbanden met ingangssignalen aangaan. De hersenen van een kind groeien bijvoorbeeld nog, en dan worden met name van dit soort verbanden nog gelegd. Al denkende leert men.

Vaak zijn signalen in de context van belang. Wie voor een zaal staat reageert bijvoorbeeld heel anders op een opmerking dan wie in diezelfde zaal zit. En op het werk gedragen we ons ook anders dan thuis. Dit soort 'context' waarin we ons bevinden helpt een neuron om te kiezen tussen wel of niet vuren; we leren ons in verschillende omgevingen anders gedragen doordat een neuron dat ons gedrag aanstuurt contact zoekt met een axon die representatief is voor deze context.

Merk op dat timing hierbij een rol speelt. Als gedrag A en B mogelijk is dan kiezen we in de ene context voor A en in de andere voor B – gewoon doordat de ingangsprikkeling in de ene context sneller leidt tot de neuronen die gedrag A uitlokken dan die voor B, en in de andere context gebeurt dat dan net andersom.

Contextgevoelige data

Het gebruik van een context om aan te geven welke uitkomsten wel en niet gewenst zijn komt ook in SQL voor. Het zijn de voorwaarden die aan een query worden gesteld, de verbanden die worden verlangd. Het grote verschil tussen een DBMS en het brein lijkt dan ook te zijn dat de condities die gebruikt kunnen worden binnen query's in ons brein zijn voorgecodeerd.

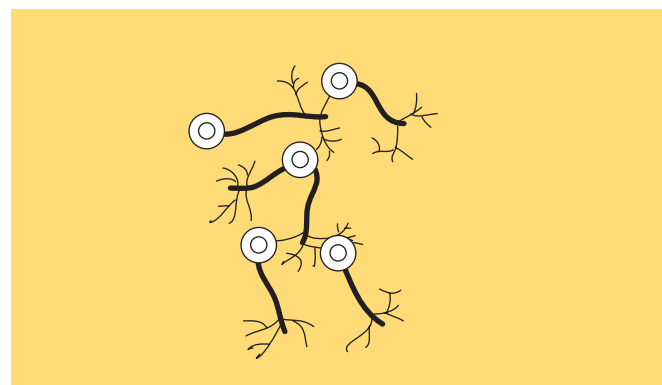
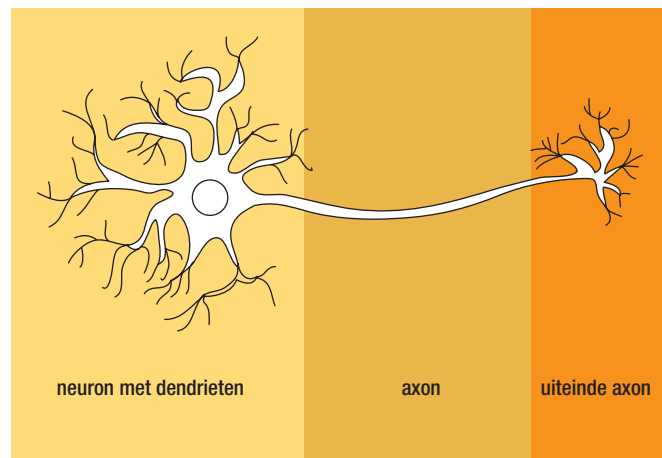
De timing van een neuron, die sneller vuurt naarmate matches

beter zijn, heeft een goede parallel in sortering van de resultaten. De dingen die meer van toepassing zijn scoren hoger (want ze arriveren eerder) dan de dingen die minder van toepassing zijn – mogelijk komen die laatste zelfs helemaal niet aan bod. Daarbij is het in neuronen wel zo dat de sorteerrichting niet kan worden omgedraaid, of er zou al een formulering van de context moeten zijn die dat tot effect heeft.

Het is interessant op te merken dat dit soort contextgevoeligheid steeds vaker opduikt als behoefte voor moderne tools en toepassingen. Een PDA die werkgerelateerde actiepunten prominenter vertoont op het werk dan thuis, is een aardig voorbeeld. Het zoeken in een plaatjesdatabase naar 'dingen die lijken op' is een ander voorbeeld.

Denken is associëren

Zoals De Bono beschrijft in *Mechanism of Mind*, kan het menselijke leerproces worden gezien als het uitslijten van paden. Naarmate we vaker een bepaald pad bewandelen, wordt het moeilijker om een ander pad er vlak naast aan te leren. Daardoor is Duits bijvoorbeeld relatief moeilijk voor Nederlanders, omdat de taal zo op de onze lijkt. Net als dat het moeilijk is om een gat in hout te boren dat deels overlapt met een eerder geboord gat. Onze hersenen vormen een enorme associërende database. Ze weten van allerlei situaties en gevallen precies wat de volgende stap moet zijn, want dat is aangeleerd. Daardoor maken we automatisch allerlei gevolgtrekkingen als er iets gebeurt. We zijn



Afbeelding 1: Schets van een neuron, en een netwerkje van neuronen.

in heel veel dingen gewoon stimulus-respons wezens. En dat houdt pas op als we dingen niet meer 'op gevoel' of 'automatisch' doen; dan pas introduceren we prikkels die de neuronen aanzetten tot andere denkpatronen en opvolgingen daarvan.

Associaties gebeuren door allerlei lagen van abstractie heen; uit onderzoek naar telegrafisten is bijvoorbeeld gebleken dat ze eerst leerden om morse te begrijpen en verzenden in de patronen van een letter tegelijk; pas later gebeurde dat in patronen van hele woorden. Opvallend in dat onderzoek is dat er een scherpe overgang plaatsvond tussen de letterfase en de woordfase; alsof er een nieuw abstractieniveau was ontstaan voor hele woorden. Veel andere dingen die we leren lijken datzelfde stramien te volgen.

Associatieve databases

Databases werken, net als ons brein, in hoge mate associatief. Een key wordt opgezocht en geeft aanleiding tot een aantal velden die verder verwerkt kunnen worden. En bij het paralleliseren van een database is een voor de hand liggende gedachte om de ruimte van alle mogelijke keys te verdelen over een aantal processoren. Het verschil zit eigenlijk in het actieve karakter van data in een brein, ten opzichte van die in een database. Want zelfs als een database verdeeld wordt over meerdere nodes, dan nog is binnen elke node sprake van een sequentieel zoekproces. In plaats van een proces van orde $O(N)$ in het moeilijkste geval zoekt een parallelle database op k nodes met een proces van de orde $O(N/k)$. Het neurale equivalent daarentegen, is van orde $O(1)$. De prijs voor deze razendsnelle opzoekmogelijkheden is natuurlijk dat er heel veel kleine processoren nodig zijn om al dat opzoekwerk te regelen. Alle data staan als het ware parallel in de aanslag, klaar om toe te springen op invoer die zoekt naar een uitkomst. Dergelijke opslagsystemen zijn niet gebruikelijk, behalve in niches zoals processor-caches. Zulke geheugens moeten omwille van de kosten relatief klein worden gehouden. Iets wat met onze hersenen kan, maar waarin database query-talen schromelijk tekort schieten, is een query in de zin van "mijmer hier eens over, zoek in deze richting en laat eens weten waar je dan op uit komt". Ofwel, het in gang zetten van associaties en het zien waar dit uit komt. Talen als SQL zijn uitermate degelijk maar laten daarmee ook een paar interessante mogelijkheden tot extrapolatie liggen.

Fouten maken is menselijk

Mensen leren van concrete situaties. Iedere volgende situatie is vaak anders, dus moet het brein gaan gokken welke reactie van toepassing is. De neuronen die het dichtst in de buurt van het voorliggende concrete geval zitten zullen het snelste vuren, zodat er meestal gewoon een reactie loskomt. Alleen heeft het brein dan wel gegeneraliseerd. En dat kan fout gaan, maar veel vaker gaat het goed.

Juist vanwege de mogelijke fouten is dit generaliserende gedrag niet wenselijk in de alledaagse toepassing van databases om bijvoorbeeld verzekeringsgegevens op te slaan, of de uitstaande boetes van een wegpiraat. Maar de veelgehoorde wens naar data-

base mining gaat wel heel sterk in de richting van generalisatie, van het vinden van verborgen verbanden in de dataset.

Zouden neurale technieken kunnen helpen om inzicht te krijgen in de verborgen verbanden in een database? Waarschijnlijk niet. Weliswaar is het denkbaar dat een neurale netwerk een database veralgemeniseert en reacties geeft voor ontbrekende delen in de dataset, maar de kennis die daartoe leidt is opgeslagen in een holistische structuur, en uit die literatuur valt evenveel af te lezen als uit koffiedik.

Fouten maken is wenselijk

Het vermogen van mensen om fouten te maken, of in elk geval om te veralgemeniseren op basis van weinig informatie, is soms erg nuttig. Met name is dit er de oorzaak van dat we creatief zijn. Een veelvoorkomende veroorzaker van creativiteit is de zogenaamde *bisociation*: een associërende denkstop die niet ontspruit uit één, maar uit twee (of meer) ingangspatronen. De twee zaken komen dan samen en combineren (of veralgemeniseren) tot een gedachte die nog niet eerder was gedacht. De ongewone combinatie leidt tot een ongewone gedachte. Veel daarvan is kaf, maar soms is zo'n ongewone gedachte ook ronduit bekoorlijk. Voor mensen bestaan technieken om creativiteit op te wekken, want creatief denken is ook maar een soort gemoedstoestand. Een voorbeeld is het in gedachten werken aan een probleem terwijl men naar tekenfilms kijkt, of door het bos wandelt zonder zich ervoor af te sluiten. Er zijn stapels van dit soort technieken om de creativiteit te prikkelen.

De cellen weten samen al de afzonderlijke stukjes informatie terug te vormen

Een database bezit deze creativiteit niet, en daar kunnen we wel blij mee zijn misschien. Wat we er wel uit kunnen leren, is dat creativiteit regelmatig absurde verbanden legt, en dat ook data mining weleens onvoorspelbare uitkomsten zal leveren als er gegeneraliseerd wordt om, bijvoorbeeld, trends of klantgedrag te voorspellen. Generaliseren over data die er niet zijn zal altijd koffiedikkijken blijven.

Google als datamijn

Een interessant onderzoek dat een beetje deze kant op gaat is het werk van Rudi Cilbrasi en Paul Vitanyi van het CWI in Amsterdam. Deze onderzoekers definieerden een 'Google Distance', die uitdrukt hoe sterk het verband tussen twee woorden is. De maatstaf daarvoor is hoe vaak Google ze gecombineerd op een pagina aantreft. Hun artikel [cs.CL/0412098](https://arxiv.org/abs/cs.CL/0412098) met de titel 'Automatic Meaning Discovery Using google' gaat diep in op de mogelijke metrieke waarmee een dergelijke woordafstand zou kunnen worden ingeschat.

Neuronen als inspiratiebron

Biologie is vaak een goede bron van inspiratie bij het oplossen van problemen in harde technische disciplines. Waar de data en de processing erover gescheiden elementen zijn in een database, bestaan de hersenen juist uit cellen die de data en verwerking daarvan samenbinden; met als gevolg een geschikter opzet voor associatief geheugen.

Het generaliserende effect lijkt nog het meest interessant voor uitbuiting in de techniek, omdat dit heel anders gebeurt dan met de normale statistische middelen. Dat kan voor data mining een nieuwe invalshoek geven. Statistiek reduceert in principe informatie tot samenvattingen, het genereert geen nieuwe informatie. Dat is tegelijk een kracht en zwakte van statistiek. De kracht daarbij is dat een indicatie kan worden gegeven van de mate van zekerheid.

Een variant op data mining is denkbaar, waarbij geen verbanden worden opgeleverd, maar waarbij á la neuraal netwerk op basis van een database wordt geëxtrapoleerd om nieuwe situaties te proberen. Het resultaat daarvan zou dan niet bestaan uit expliciete verbanden, maar op responses op een experiment, een balletje dat wordt opgegooid: hoe zou men reageren als we X zouden doen? Ook hier bestaat een indicatie voor de zekerheid in de vorm van de vertragingstijd van de reactie, maar die maatstaf is veel minder betrouwbaar dan hetgene de statistiek oplevert.

Er lijken dus wel degelijk mogelijkheden te bestaan om op basis van de in dit artikel getrokken parallel nieuwe technieken uit te proberen, maar het gaat in ieder geval ten koste van de zekerheid die we kennen van andere toepassingen van databases. En dat is lang niet overal wenselijk.

Conclusie

Hersenen slaan informatie op, en weten die op miraculeuze wijze te reconstrueren. De manier waarop dat gebeurt is heel anders dan in een DBMS, en voor een deel zijn daar interessante verbanden te leggen. Toch zijn vooral de verschillen kenmerkend, met name zorgt de generaliserende neiging van het brein voor inaccurate en daardoor soms zelfs creatieve responses.

Literatuur

Rudi Cilibrasa, Paul Vitanyi. *Automatic Meaning Discovery Using Google*, www.cwi.nl/~paulv/papers/amdug.pdf

John Hertz, Anders Krogh, Richard Palmer. *Introduction to the Theory of Neural Computation*, ISBN 0-201-51560-1

Edward De Bono. *The Mechanism of Mind*, 1968.

Edward De Bono. *Lateral Thinking, Creativity Step by Step*.

Rick van Rein

Dr. ir. H. van Rein (rick@openfortress.nl) is ontwikkelaar en beheerder bij OpenFortress Digital signatures.

DE ICT-PARTNER

voor uw comfort



Gelijk water is uw organisatie continu in beweging. Groeien, reorganiseren, exporteren, fuseren, privatiseren. Inspelen op ontwikkelingen. Ontdekken van nieuwe diensten en markten. U wilt vooruit en ook controle houden. De juiste beslissingen nemen. Daarbij vertrouwen op uw ICT-organisatie. Goed gestructureerd, ingericht met beleid, voorzien van betrouwbare verbindingen. Kies dan voor Solipsis. Jarenlange ervaring maakt ICT onze tweede natuur en ons uw betrouwbare en stabiele partner. *Wel zo comfortabel.*

SOLIPSIS
COMFORT CLASS IN ICT

www.solipsis.nl

telefoon: (0418) 57 61 00 info@solipsis.nl