WOORDEN WIKKEN EN WEGEN

Woorden wikken en wegen

Rede uitgesproken bij de aanvaarding van het ambt van hoogleraar Information filtering and aggregation aan de Faculteit der Natuurwetenschappen, Wiskunde en Informatica van de Radboud Universiteit Nijmegen op donderdag 25 juni 2009

door prof. dr. ir. Wessel Kraaij

Vormgeving en opmaak: Nies en Partners bno, Nijmegen
Fotografie omslag: Bert Beelen
Drukwerk: Drukkerij Roos en Roos, Arnhem

Woorden wikken en wegen. Foto weegschaal: Josselien van Eijk

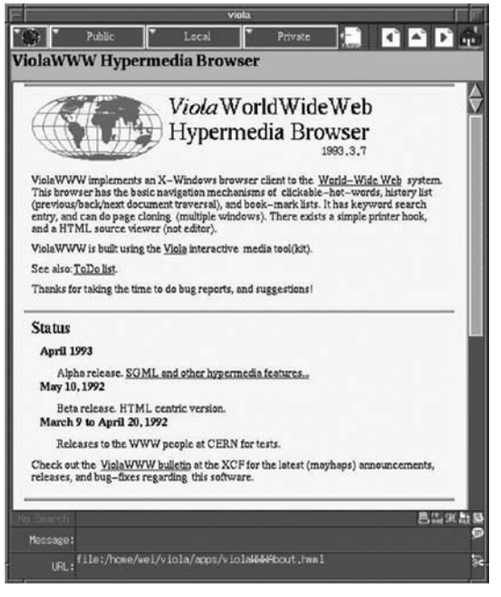
ISBN 978-90-9024448-8

Mijnheer de rector magnificus, geachte collega's, familie en vrienden, dames en heren.

Zoals eenieder zich nog wel de roes kan herinneren na de eindexamenuitslag van de middelbare school, zo heeft voor mij ook de eerste kennismaking met het world wide web een bijzondere indruk gemaakt. Het zal ergens in 1993 geweest zijn toen ik aan collega's van het toenmalige Instituut voor Taal en Kennistechnologie van de Universiteit van Tilburg de Viola[1] web browser liet zien op de enorme monitor van mijn Digital VAX workstation. De opwinding om zo maar real time van het ene naar het andere continent te kunnen springen met muisklikken en met foto's opgemaakte webpagina's te kunnen bekijken.

Wat nu triviaal klinkt was destijds revolutionair. Het web was nog overzichtelijk, zo waren er sites met landkaartjes waarop je kon zien in welke stad er een website aanwezig was. Door op een stipje te klikken kon je direct doorklikken naar de site, allemaal volgens een gestandaardiseerd open protocol.

In 1993 werden computernetwerken al op beperkte schaal gebruikt in de universitaire gemeenschap voor e-mail, discussiefora[2] en de distributie van gegevens, ook bestond er al een hypertext community[3]. Maar de koppeling van hypertext met netwerktechnologie, in combinatie met een grafische browser, betekende een revolutie in het toegankelijker maken van informatie.



Figuur 1:
Viola web browser

1. INLEIDING

De samenleving is door de komst van internet en het web ingrijpend veranderd en verandert nog steeds. De globalisering is erdoor in een stroomversnelling geraakt[4] en oude kaders moeten worden herzien. Dit zien we bijvoorbeeld in de media, maar ook in de communicatie, dienstverlening, detailhandel, productinnovatie et cetera.

Een aantal voorbeelden: de vaccinatiecampagne van het RIVM ter voorkoming van baarmoederhalskanker is mislukt omdat de impact van verhalen die de ronde doen op het internet is onderschat. De meiden van nu communiceren veel via chat en sociale netwerksites en worden minder effectief bereikt via klassieke overheidscampagnes.

Het web 2.0, dat het voor iedereen gemakkelijk heeft gemaakt zijn mening, foto's of filmpjes met een groot deel van de wereldbevolking te delen, leidt tot meer transparantie. Iedere groepering heeft zo zijn eigen podium gevonden, we kunnen nu meelezen met weblogs van soldaten in Afghanistan, wetenschappers en politici. President Obama heeft tijdens zijn campagne laten zien dat je met web 2.0-technieken een beslissende invloed kunt realiseren.

2. UITDAGINGEN VAN DE INFORMATIEMAATSCHAPPIJ

De huidige informatiemaatschappij stelt zowel de gebruiker als de aanbieder van informatie voor nieuwe uitdagingen. Hoe kunnen we omgaan met de enorme dagelijkse aanwas aan nieuwe informatie?[5]

Het vinden van de juiste informatie

Kennis is macht[6] luidt een bekend aforisme van Francis Bacon waarmee een machtsverhouding wordt geïmpliceerd tussen kennishouders en degenen die geen directe toegang hebben tot kennis. De toegang tot informatie is nog nooit zo laagdrempelig geweest. Het beschikbaar komen van relevante informatie kan een grote emanciperende kracht hebben, denk aan de akkerbouwers in Afrika die door de komst van mobieltjes opeens de actuele wereldmarktprijs kennen van hun producten. Maar doordat een centrale kwaliteitscontrole van informatie op het internet ontbreekt, is toegang tot informatie alleen niet voldoende voor kennis en dus ook niet voor macht. In deze door media en communicatie gedomineerde samenleving worden mensen voortdurend ondergedompeld[7] in informatie. Om in deze huidige maatschappij succesvol te opereren is het daarom belangrijk om de juiste informatiebronnen te selecteren, te combineren en in context te plaatsen. Anders gezegd: de juiste informatie moet op het juiste moment op het juiste abstractieniveau beschikbaar zijn. Technologie die gebruikers en aanbieders daarmee ondersteunt is van vitaal belang voor onze kenniseconomie.

Figuur 2: Vir sapiens fortis est nam et ipsa scientia potestas est (Francis Bacon)

De huidige search engines schieten vaak nog sterk tekort doordat ze niet zijn ontworpen vanuit het besef dat informatie vergaren een stapje vormt van het proces naar een concreet doel van de eindgebruiker. Daarnaast beseffen eindgebruikers ook meestal niet dat het moeilijk is om een zoekvraag van twee woorden effectief te honoreren zonder begrip van de context. Idealiter zou een search engine moeten weten welke informatie ik al ken, en in welke onderwerpen ik geïnteresseerd ben. Op die manier zou alleen nieuwe relevante informatie worden gepresenteerd en kan de juiste interpretatie worden gekozen van ambigue zoekvragen. Bijvoorbeeld het eiland Java, als ik een reis aan het plannen ben, of de insectenpagina als mijn zoontje een spreekbeurt wil voorbereiden en zoekt op 'kever'. Nu wordt in deze gevallen aangesloten bij de interpretatie van de meerderheid van de gebruikers, in dit geval de programmeertaal en een populaire oldtimer.

Daarnaast zou het wenselijk zijn om een aggregatieslag te maken over de enorme lijst van zoekresultaten. *Wat zijn de hoofdzaken? Kan vrijwel identieke en dus overbodige informatie worden onderdrukt?* Over sommige onderwerpen zou ik slechts op hoofdlijnen willen worden geïnformeerd, op andere thema's zou ik alle details willen volgen. Als de hoeveelheid informatie en de aard en tijdstip van presentatie helemaal zouden kunnen worden gepersonaliseerd, zou 'informatiestress'[8] ten gevolge van de druk om 'bij te blijven' mogelijk kunnen worden voorkomen.

Het beoordelen van de betrouwbaarheid van informatie

Een probleem van fundamentele aard is de selectie en duiding van informatie die via het internet wordt gevonden. Welke informatie is gevalideerd? Welke bronnen zijn betrouwbaar? Hoe om te gaan met tegenstrijdige informatie?

Ook vóór de opkomst van internet was het gewicht van informatie al gekoppeld aan het gezag van de bron. Iemand heeft gezag als zijn of haar mening gevolgd wordt. De woorden van een gezaghebbend persoon hebben dus een hoger gewicht dan die van iemand met een lagere autoriteit. Het draait in het maatschappelijk verkeer voor een belangrijk deel om invloed en het verwerven daarvan. De autoriteit van een bron was vroeger voor een deel gekoppeld aan sociale status, misschien ook omdat de toegang tot kwaliteitsinformatie geprivilegieerd was. Met de komst van internet is er meer dynamiek ontstaan in het landschap van bronnen met gezag. Door een goede weblog te beginnen, kun je in relatief korte tijd een expertstatus op een bepaald terrein verwerven. Het staat nog te bezien in hoeverre weblogs ook echt als gezaghebbend kunnen worden gekwalificeerd, in de zin dat mensen de informatie gebruiken als leidraad voor belangrijke beslissingen. Het schatten van de betrouwbaarheid van informatie is van groot belang voor journalistiek, opsporing en inlichtingenwerk, maar is ook een opgave waarmee iedereen die internet gebruikt dagelijks geconfronteerd wordt. *Zijn de reviews van deze digitale camera authentiek? Wat betekent het dat dit hotel louter positieve reviews heeft? Is de informatie over een nieuwe medicatie of behandelmethode betrouwbaar?*

De bescherming en benutting van persoonsgebonden informatie

Tot nu toe heb ik vooral gesproken over het perspectief van de gebruiker. De evolutie van search engines heeft geleid tot een businessmodel waarin aanbieders van informatie geen vergoeding ontvangen en ook gebruikers niet hoeven te betalen. Het gaat allemaal om advertenties, en search engines hebben belang bij een zo nauwkeurig mogelijk beeld van de intenties van de gebruiker om gerichte advertenties te plaatsen.

Dat heeft geleid tot een personalisatiearchitectuur die centraal is georganiseerd en vanwege de daaraan gekoppelde privacyproblemen nooit echt populair geworden is. Ook data mining op persoonsgerelateerde gegevens bevindt zich in dit spanningsveld. Data mining wordt onder andere toegepast voor marketing, fraudedetectie, het bepalen van risicoprofielen in de verzekeringswereld of het detecteren van cybercrime.

Net zoals we spreken over de preventieve en curatieve gezondheidszorg, kan men spreken over actieve en reactieve toepassingen van data mining. Bij actieve toepassingen wordt het gedrag van gebruikers online gemonitord of geclassificeerd om tijdig te kunnen interveniëren. Bij reactieve toepassingen gaat het meer om forensisch analytisch onderzoek nadat een bepaald (meestal ongewenst) voorval heeft plaatsgevonden. Vooral het actief vergaren van informatie over gedrag wordt doet veel mensen aan 'big brother' denken. Maar ook de wettelijke regelingen met betrekking tot dataretentie van internet service providers[9] ten behoeve van forensisch onderzoek roepen vaak heftige reacties op, omdat er geen zicht is op wat er met de gegevens gebeurt en wie daar toegang toe hebben.

Hier ook weer een spanningsveld tussen de belangen van het individu en die van een bedrijf of de samenleving als geheel. Een groot deel van de jonge generatie lijkt zich niet echt zorgen te maken en profileert zich volop op sociale netwerksites op zoek naar nieuwe contacten en kansen. Dat heeft misschien ook te maken met het feit dat hun leven en netwerk nog in opbouw zijn en de voordelen van selectieve transparantie groter zijn dan de nadelen.

3. ONDERZOEKSAGENDA

Ik heb met een aantal voorbeelden geïllustreerd dat internet een diepgaande invloed heeft op de manier waarop we communiceren, studeren, ons werk doen, en ons ontspannen. Het internet biedt vele mogelijkheden omdat het gemakkelijk is om nieuwe verbindingen te leggen of nieuwe trends te ontdekken, maar is ook een bedreiging omdat privacygevoelige informatie gemakkelijk in verkeerde handen kan vallen, communicatiekanalen dreigen te verstoppen door overbelasting en omdat het niet meer zo gemakkelijk is om de betrouwbaarheid van informatie te beoordelen.

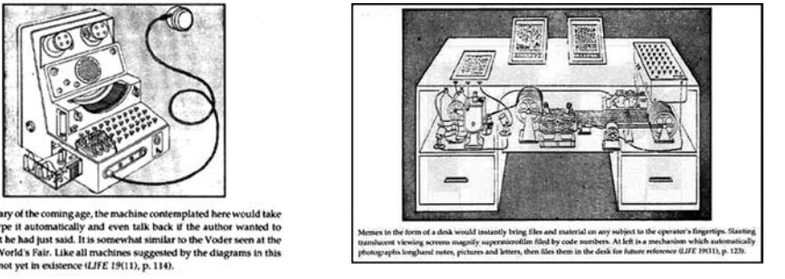
Ik wil in deze rede niet zozeer terugblikken op mijn eigen recente onderzoek maar vooral aandacht besteden aan drie onderzoeksthema's die ik onder de vlag van het Information Foraging Lab van de Radboud Universiteit in de komende jaren verder wil uitwerken met als doel het proces van informatie verzamelen, beoordelen en verwerken te verbeteren. De eerste twee thema's zijn gekoppeld aan de twee centrale uitdagingen die ik zojuist heb ingeleid. Als eerste zal ik ingaan op het thema 'verbeterde relevantie door impliciete gebruikersmodellering' dat gekoppeld is aan de uitdaging om de juiste informatie te vinden. Het tweede thema richt zich op het bepalen van 'reputatie en betrouwbaarheid'. Het derde thema betreft de 'kwantitatieve evaluatie van privacyaspecten van informatiesystemen'. Deze drie thema's sluiten goed aan bij de onderzoeksagenda ICT2030 van ICT regie[10], in het bijzonder bij de thema's 'beheersen van complexiteit' en 'vergroten (van) vertrouwen'.

4. THEMA 1: VERBETERDE RELEVANTIE DOOR IMPLICIETE GEBRUIKERSMODELLERING

Informatiewetenschap = informatietheorie & gedragswetenschap

Ik wil nu kort terugblikken op de geschiedenis van de geautomatiseerde analyse van taal en tekst, omdat het laat zien dat de computationele taalkunde en information retrieval dezelfde wortels hebben, decennia lang naast elkaar zijn geëvolueerd en nu weer naar elkaar toe groeien.

Toegang tot informatie door digitalisering was al een droom van Amerikaanse denkers in de jaren veertig van de vorige eeuw. In het beroemde essay 'As we may think' dat Vannevar Bush, de directeur van het Amerikaanse overheidsbureau voor wetenschappelijk onderzoek, schreef direct na de afloop van de Tweede Wereldoorlog[11], wordt een pleidooi gehouden om de natuurkunde weer in te zetten voor vreedzame doeleinden. In deze visionaire tekst heeft Bush zich laten inspireren door het menselijke cognitieve systeem. Hij stelt voor om de menselijke kennis die door de eeuwen is opgebouwd te ontsluiten via een associatief model dat geïmplementeerd is als een utopisch systeem,



Supersecretary of the coming age, the machine contemplated here would take dictation, type it automatically and even talk back if the author wanted to review what he had just said. It is somewhat similar to the Voder seen at the New York World's Fair. Like all machines suggested by the diagrams in this article, it is not yet in existence (*LIFE* 19(11), p. 114).

Memex in the form of a desk would instantly bring files and material on any subject to the operator's fingertips. Slanting translucent viewing screens magnify supermicrofilm filed by code numbers. At left is a mechanism which automatically photographs longhand notes, pictures and letters, then files them in the desk for future reference (*LIFE* 19(11), p. 123).

Figuur 3a eb 3b: (evt onderschrift splitsen) Twee illustraties van de droombeelden van Vannevar Bush. Links een prototype dictafoon, waarin een microfoon aan een elektrische typemachine is gekoppeld, rechts de Memex.

de ‘Memex’, voorzien van een spraaktranscriptiesysteem, optische scanner, microfilmgeheugen en mogelijk zelfs rechtstreekse aansturing vanuit de menselijke hersenen. Bush zelf beschouwde het vastleggen van associatieve gedachtesporen in de vorm van annotaties die twee informatiebronnen koppelen als zijn meest fundamentele idee. Een dergelijke associatieve structuur is inmiddels gemeengoed geworden door de hypertextstructuur van het web. In de jaren dertig was Bush decaan bij het MIT. Een van zijn studenten was Claude Shannon, de grondlegger van de informatietheorie. Shannons werk[12] heeft grote invloed gehad op de ontwikkeling van taal- en spraaktechnologie, onder andere op automatisch vertalen.

Een andere pionier op dit vlak was Warren Weaver, die in 1949 in een memorandum[13] (tegenwoordig zou dat misschien een weblog-posting zijn) vier mogelijke ‘aanvalsroutes’ schetst met betrekking tot automatisch vertalen. De eerste gaat uit van het feit dat woordbetekenis voor een belangrijk deel wordt vastgelegd door de onmiddellijke context. De tweede aanpak volgt de lijnen van de formele logica. De derde aanpak is gestoeld op de stochastische informatietheorie van Shannon. De laatste, meest ambitieuze aanpak veronderstelt de ontdekking of ontwikkeling van een generieke tussentaal. Zestig jaar later blijken vooral de statistische modellen succesvol. We weten nu dat geautomatiseerde semantische analyse van natuurlijke taal eigenlijk alleen haalbaar is binnen een zeer beperkt domein. De methoden en modellen zijn eenvoudigweg niet robuust genoeg voor de interpretatie van korte snippers van informatie, die wemelen van de semantische en syntactische ambiguïteiten. Het menselijk interpretatie- en adaptatievermogen is op dit vlak nog ongeëvenaard.

De problematiek van automatisch vertalen is uiterst relevant voor het vakgebied information retrieval. Het gaat daar immers ook om het leggen van een verbinding tussen de terminologie van degene die zoekt en de terminologie van de informatieaanbieder. De information retrieval heeft zich in de jaren vijftig van de vorige eeuw ontwikkeld als een taalkunde-arme manier om informatie te ontsluiten. Het zogenaamde ‘bag of words’-model[14], waarin tekst wordt gereduceerd tot een verzameling ongeordende woorden is leidend geweest. Dit model is in zekere zin rechtstreeks terug te voeren op het eerste model van Weaver. Door de beschikbaarheid van steeds meer data, kunnen kleine stapjes gezet worden naar verbeterde, meer complexe modellen. In de information retrieval wordt de theorievorming gestuurd door empirische resultaten en andersom. Een methodiek die voor het eerst werd gepropageerd door de eerder genoemde Francis Bacon.

In mijn eigen bijdragen aan information retrieval-onderzoek in de afgelopen vijftien jaar heb ik gewerkt in de traditie van de generatieve taalmodellen van Shannon. Een taalmodel kan bijvoorbeeld worden voorgesteld door een vaas met stukjes papier. Op ieder papier is een woord gedrukt. Stelt u zich even voor dat we een boek met een klein schaartje helemaal verknippen en een vaas vullen met de woordjes. Nu kan nieuwe tekst worden gemaakt door steeds een papiertje te pakken, het over te schrijven en het

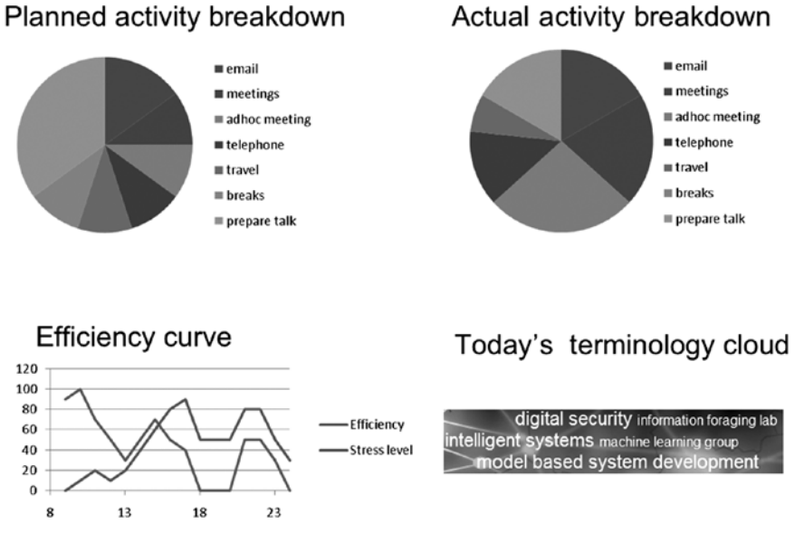
Figuur 4: Generatief unigram taalmodel, ieder woord wordt onafhankelijk gekozen van de voorgaande woorden. (Bron: Floddertje, A.M.G. Schmidt)

weer terug te doen in de vaas. De kans op een bepaald woord wordt gewogen door het relatieve aantal voorkomens van dat woord in het oorspronkelijke boek.

Dit model van gewogen woorden is zo eenvoudig dat het ruim toepasbaar is. Je zou het kunnen zien als een *lingua franca* in de zin van Weaver, omdat er eenvoudige, goed gedefinieerde metrieken bestaan om de afstand tussen twee taalmodellen te bepalen. Wel geldt dat hoe minder context er beschikbaar is, hoe minder onderscheidend het taalmodel is.

Het is interessant om een aantal van de grote successen van information retrieval-onderzoek eens nader te beschouwen. In de jaren zeventig van de vorige eeuw werd overtuigend aangetoond dat ‘relevance feedback’ – dat is het interactief waarderen van zoekresultaten door de gebruiker en de terugkoppeling daarvan naar het systeem – tot een grote verbetering van zoekresultaten leidt. In de jaren negentig bleek dat de resultaten van web search engines structureel konden worden verbeterd door te kijken naar hyperlinks. Hoe meer links naar een pagina, des te belangrijker deze is. Bovendien is de zogenaamde *anchor*-tekst vaak zeer goed bruikbaar als additionele index-descriptor voor een pagina. In het afgelopen decennium is die trend verder doorgezet, search engines maken nu op grote schaal gebruik van het zoekgedrag van alle gebruikers van de zoekdienst. Men zou kunnen concluderen dat door mensen geproduceerde metadata in de vorm van expliciete of impliciete waarderingen of semantische annotaties het grootste potentieel vormen voor de verbetering van zoektechnologie.

Als we het privacyaspect en de advertentiebelangen van de grote search engine-bedrijven compleet buiten beschouwing zouden laten, is het dus lonend om alle sporen van interactie van gebruikers met informatie te volgen en in kaart te brengen. Op die manier kan er maximaal geprofiteerd worden van de collectieve toegevoegde waarde. Uit de enorme hoeveelheid klikinformatie kunnen nieuwe associaties worden gedestilleerd. Een beeld dat ook past in de visie rond de Memex. Er zijn aanwijzingen dat ook met een persoonlijke decentrale en dus meer privacyvriendelijke variant van context en activiteitmodellering nog veel winst te behalen valt. We zien dus dat het information retrieval-domein steeds meer opschuift in de richting van de gedragswetenschappen. Door de interactie van gebruikers met informatie te modelleren en in de context van taken en doelen te plaatsen[15], is er waarschijnlijk nog een grotere verbetering te boeken. Gebruikers nemen tijdens de interactie met search engines voortdurend beslissingen om zo optimaal mogelijk resultaat te boeken: zoek ik nog even door of pas ik de zoekvraag aan of besluit ik via een andere weg verder te zoeken? Het theoretische kader van information foraging van Peter Pirolli waarin een analogie beschreven wordt tussen de sociale activiteit van



Figuur 5: Schets van een persoonlijk informatie- en activiteitendashboard.

voedsel verzamelen bij dieren en het vergaren van informatie, biedt op dit punt een opening om vanuit de strikt mathematische modellering van informatie een brug te slaan naar het modelleren van menselijk zoekgedrag.

In het onderzoek van mijn groep zal ik aandacht gaan besteden aan het monitoren van het individuele gedrag van kenniswerkers, door hun communicatie, computer- en internet-interactie automatisch te analyseren in termen van interesse, expertise, tempo en tijdsbesteding. We willen ook onderzoeken of deze feedbackgegevens kunnen worden ingezet als een soort persoonlijke coach. Daarbij kunnen sensoren worden ingezet om bijvoorbeeld stressniveaus te meten. De hypothese is dat werknemers op deze manier beter inzicht kunnen krijgen in hun manier van werken. Het is de bedoeling om ook te experimenteren met verschillende methoden en architecturen voor het bijhouden van een publiek profiel op basis van het afgeschermde privéprofiel. Met taalmodellen kunnen dan automatisch associaties worden gelegd. Op die manier kunnen mensen automatisch worden geattendeerd op een nieuwe collega, of een voor hen interessant artikel. Een goed persoonlijk profiel zou als het ware de juiste informatie kunnen aantrekken. Het proces van actief informatie zoeken verschuift zo naar het passief ontvangen van gepersonaliseerde informatie.

5. THEMA 2: MODELLEN VOOR REPUTATIE EN BETROUWBAARHEID

het in kaart brengen van communities

Ik sprak al eerder over het belang van vertrouwen: vertrouwen in mensen, in informatie en in de beveiliging van informatiesystemen. Het web is een zeer laagdrempelig publicatiekanaal, waar kwaliteitsinformatie moeilijker te vinden is. Het web is ook een momentopname. We kunnen meestal niet de vorige versies van een webpagina bekijken of hun herkomst bepalen[16]. Er is daarom grote behoefte aan indicatoren voor de betrouwbaarheid van informatie, bijvoorbeeld rond medische vragen maar ook rond aanbevelingen uit sociale netwerken.

Vertrouwen is een thema dat vanuit verschillende wetenschappelijke disciplines wordt bestudeerd, bijvoorbeeld vanuit de sociale wetenschappen. In de informatie- en communicatietechnologie is betrouwbaarheid nauw verbonden met security, maar betrouwbaarheid is ook een centraal onderdeel van de visie op het semantische web[17]. Ik wil me nu beperken tot de betrouwbaarheid van informatie. In een recente studie[18] noemen Gil en Artz maar liefst negentien factoren die ons oordeel over de betrouwbaarheid van informatiebronnen op het web beïnvloeden. Enkele voorbeelden: de reputatie van een site, auteur of organisatie, aanbevelingen van derden, directe ervaring in het verleden, verwijzingen vanuit andere bronnen, spelling, grammatica, opmaak van de site, populariteit, publicatiedatum. De reputatie van entiteiten die gekoppeld zijn aan de informatie is van groot belang, en te berekenen bijvoorbeeld door middel van citatiescores. Het creëren van een robuust model voor betrouwbaarheid is niet gemakkelijk. Veel indicatoren zijn slechts met heuristieken te schatten en er is bovendien inter-

actie tussen de verschillende parameters. Een over het algemeen betrouwbaar iemand kan niet over alle onderwerpen goede aanbevelingen doen. Andersom kan een minder betrouwbaar iemand op een deelterrein misschien wel een heel goede aanbeveling doen.

In het onderzoek gekoppeld aan mijn leerstoel wil ik mij vooral richten op het modelleren van de dynamiek en sociale structuur van web communities. Een veelbelovende aanpak waar nu al door TNO-collega Stephan Raaijmakers aan gewerkt wordt, is het in kaart brengen van het ecosysteem van sociale netwerksites. De hypothese is dat we door naar het dynamische gedrag van dergelijke sites te kijken, inzicht kunnen krijgen in de sociale structuur van de groep, met in het bijzonder aandacht voor aspecten zoals autoriteit, betrouwbaarheid, en de snelheid van informatiedisseminatie. Een eerste resultaat is een model voor de disseminatie van informatie gebaseerd op thermodynamische modellen, gecombineerd met de analyse van wederzijdse feedback van groepsleden door het toepassen van sentimentanalyse. Op deze manier kan de relatieve autoriteit van verschillende leden van het sociale netwerk worden berekend. Dit kan helpen om de aanbeveling van onbekende personen op waarde te schatten. Een volgende stap is de verfijning van het model met een semantische component of parameters zoals

Figuur 6: Door de analyse van sociale netwerken kunnen verschillende rollen worden geïdentificeerd (illustratie: http://hci.stanford.edu/jheer/projects/vizster/).

geografische afstand of smaak. Een andere mogelijke invalshoek om grip te krijgen op betrouwbaarheid is om de hyperlinkstructuur van het web te verrijken met type-informatie die aangeeft of een verwijzing positief, neutraal of negatief is[19]. Het PageRank-algoritme van Brin & Page[20] gaat immers uit van louter positieve links. De polariteit van links zou kunnen worden uitgerekend door de lokale context van de links te analyseren.

Om dergelijke technieken op grote datasets toe te kunnen passen, is het niet realistisch om de complete webgraaf of sociale graaf bekend te veronderstellen. Er wordt daarom hier aan de Radboud Universiteit gewerkt aan technieken die het gewenste resultaat benaderen door zich tot een relatief kleine context te beperken.

6. THEMA 3: EVALUATIE VAN PRIVACYASPECTEN

naar een persoonlijke geïnformeerde afweging tussen kansen en risico's

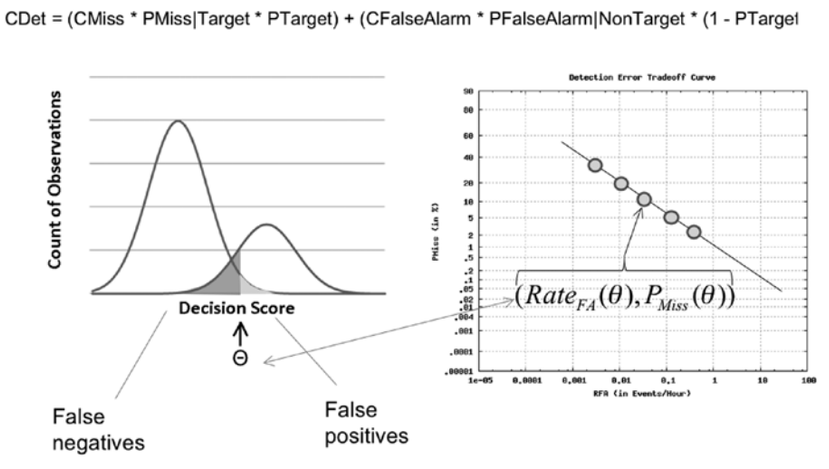
Ik wil eindigen met een thema dat me na aan het hart ligt, namelijk het ontwerp van evaluatiemethodieken. In de twee thema's die ik besproken heb, heb ik betoogd dat informatietoegang kan worden verbeterd door te kijken naar de interactie van individuen met informatie en met elkaar. Op basis van die interactie kunnen namelijk profielen worden gemaakt die helpen bij gepersonaliseerd zoeken, het koppelen van de juiste informatie of personen aan iemands profiel, en het mogelijk maken om de betrouwbaarheid van informatie of autoriteit van een bron te kunnen bepalen. Tot nu toe wordt dit type gedetailleerde persoonsprofilering niet of nauwelijks toegepast, omdat het exploiteren of beschikbaar maken van dergelijke informatie in een centrale database zo gevoelig ligt. Ook voor wetenschappelijk onderzoek is dit type data maar zeer beperkt toegankelijk. Dat neemt niet weg dat er wel decentrale oplossingen denkbaar zijn, zoals personalisatie via een proxy op de eigen pc van individuele gebruikers[21].

In het publieke debat over de toenemende digitalisering van persoonsgebonden informatie zijn de meningen sterk verdeeld. Tegenstanders benadrukken de mogelijke schade, voorstanders benoemen de successen van data mining en wijzen op de meer laconieke houding van de jongere generatie ten aanzien van persoonlijke informatie. Om het debat goed te kunnen voeren is enige nuancering gewenst.

Om te beginnen is het goed om ons te realiseren dat niet alle persoonsgebonden informatie even privacygevoelig is en dat de afweging welke informatie gedeeld wordt en welke informatie privé blijft per individu sterk kan verschillen. Het delen van informatie via een publiek persoonlijk profiel kan, zoals gezegd, ook voordelen hebben. Transparantie kan een belangrijke bijdrage leveren aan vertrouwen. Het al of niet aan derden ter beschikking stellen van een persoonlijk profiel is dus te zien als een proces van wikken en wegen. Hoe valt de balans uit? Weegt het potentiële nut zwaarder of zijn de vermeende risico's groter? Het zou aantrekkelijk zijn als die kansen en risico's ook op een bepaalde manier kunnen worden gekwantificeerd. Stel dat we de mogelijke schade die een persoon kan worden berokkend door openbaarmaking van persoonsgebonden

gegevens kunnen schatten, en dat we ook weten hoe groot de waarschijnlijkheid op ongeoorloofd gebruik van privégegevens is. Bij schade kun je denken aan financiële schade, reputatieschade, maar ook – in bijzondere gevallen - psychische en/of fysieke schade. Hoe groter de potentiële schade, hoe gevoeliger de informatie.

Het uitruilen van verschillende belangen of, anders geformuleerd, het maximaliseren van de systeemprestatie gegeven twee omgekeerd gecorreleerde indicatoren, is ook een bekend gegeven in de information retrieval. Zo worden detectiealgoritmen vaak geëvalueerd aan de hand van een zogenaamde kostencurve, waarin gegeven een bepaalde instelling van het algoritme en geparametriseerde kosten van de verschillende typen fouten een totaal kostenniveau wordt berekend. Door de instelling van het systeem aan te passen, verandert de verhouding tussen de twee typen foutenmarges en daarmee veranderen de totale kosten. Ook in de medische wereld zijn er indicatoren die de kwalitatieve aspecten van medisch handelen kwantificeren in termen van gewonnen levenskwaliteit (QALY: Quality averaged life years) of gereduceerde kwaliteitsvermindering (DALY: disability adjusted life years)[22]. De parameters in deze modellen worden geschat door interviews. Door een dergelijke kwantitatieve kosten-batenanalyse kunnen moeilijke ethische keuzes ten aanzien van de inzet van beperkte middelen enigszins worden geobjectiveerd. Het gaat net als bij de privacydiscussie om de afweging van een maatschappelijk belang ten opzichte van het belang van individuen.



Figuur 7: Voorbeeld van de berekening van een kostencurve.

Als gedachte-experiment stel ik voor om een privacybeschermend systeem te evalueren met een kostenmodel[23]. Het eenvoudigste is om dit systeem te zien als een poortwachter die checkt of de bedoelingen van de informatievrager legitiem zijn. We kunnen de performance van een dergelijk systeem simuleren door te kijken naar situaties waarin legitieme vragen veel vaker voorkomen dan niet legitieme, of andersom. De kosten van het systeem vallen uiteen in de zogenaamde *false positives,* de gevallen waarin ten onrechte toegang gegeven wordt tot de gegevens en de *false negatives*, de gevallen waarin ten onrechte geen toegang wordt gegeven tot de data. De totale kosten zijn dus een som van opgelopen schade en de kosten geassocieerd met gemiste kansen. Elke kostencomponent is het product van de kans dat het systeem een fout maakt gegeven een bepaalde context, de kans op deze context en de relatieve kosten die met de gebeurtenis verbonden zijn. Om een dergelijk kwantitatief model te kunnen toepassen zijn schattingen van de kans op het aantal voorkomens nodig en parameters die de 'kosten' en het 'nut' van voordelen en nadelen kwantificeren en uiteraard metingen aan het detectiesysteem zelf. Het eerste type parameters kan geschat worden door te kijken naar ervaringsgegevens. Het tweede type parameters vraagt om een ethische en politieke afweging.

Het aantrekkelijke aspect van de kostencurves zit hem in het feit dat de kosten kunnen worden geminimaliseerd door een parameter in het systeem aan te passen en het feit dat de afweging transparant is. Om een dergelijke afweging ook in het data-mining-versus-privacy-debat te kunnen maken zou het mogelijk moeten kunnen zijn om met een 'regelknop' persoonlijke informatie meer of minder makkelijk toegankelijk te maken. In een ideale wereld zou het betrokken individu zelf een dergelijke instelling moeten kunnen maken, daarin gesteund door het model. Een continue regelknop is niet direct te realiseren, maar kan wel benaderd worden door discrete niveaus. Het zou nuttig zijn om vooraf de gevoeligheid van verschillende typen van persoonsgebonden informatie te bepalen, bijvoorbeeld door het toekennen van verschillende niveaus van vertrouwelijkheid. Voor ieder niveau kan dan een adequate en proportionele beschermingsprocedure worden geïmplementeerd die aangeeft wie er toegang heeft tot de betreffende informatie en onder welke voorwaarden.

Bij veel toepassingen in de marketing- en dienstencontext is het vaak niet eens nodig om gedragsgegevens te kunnen koppelen aan individuen en volstaan geanonimiseerde geaggregeerde gegevens voor het optimaliseren van diensten of opsporingsdoeleinden[24]. Enkele voorbeelden: de intelligente energiemeter heeft een interessant potentieel voor energiebesparing: als een huishouden toegang heeft tot de eigen informatie, kan er een zeker feedbackmechanisme op gang worden gebracht[25]. De betrokken energiemaatschappij kan de variatie in capaciteitsvraag ook prima meten met een meter in de wijkcentrale. Op die manier wordt misbruik van gegevens voorkomen. In een forensische context is het uiteindelijk wel wenselijk om informatie uit communicatieverkeer te koppelen met de identiteit van personen. Zo werd er dit voorjaar een jongeman aangehouden[26] die op de Amerikaanse website 4chan.org had aangekondigd een bloedbad

aan te richten op een school in Nederland. De identificatie van de dader was mogelijk doordat details over internetverkeer op deze Amerikaanse website gekoppeld konden worden aan naam en adresgegevens via de administratie van een Nederlandse internet service provider. De dader bleek overigens gebruik te maken van een onbeveiligd draadloos netwerk van de buren, gezond verstand blijft dus nodig.

Voor de oplossing van het privacydilemma is het daarom belangrijk dat er vanuit publieke kennisinstellingen alternatieve, transparante informatiearchitecturen worden ontwikkeld, waarin de belangen van individuele burgers versus marktpartijen en/of overheid zorgvuldig worden geïncorporeerd. Een kwantitatief evaluatiemodel is daarbij van groot belang om de complexe afweging tussen kansen en risico's zoveel mogelijk te objectiveren.

Ik heb laten zien dat het nodig is om expertise uit verschillende wetenschappelijke disciplines te combineren om de uitdagingen van de informatiemaatschappij aan te pakken. In de uitvoering van deze onderzoeksagenda wil ik daarom nadrukkelijk de samenwerking opzoeken met andere faculteiten, verwante onderzoeksgroepen in Nederland en ook daarbuiten. Een begin is gemaakt door de oprichting van het Information Foraging Lab, waarin op het vlak van onderzoek en onderwijs wordt samengewerkt met de taal- en spraaktechnologen van de Faculteit der Letteren, maar van waaruit in de toekomst ook samenwerking zal worden gezocht met gedragswetenschappers. De ambitie is het uitvoeren van transdisciplinair onderzoek, waarin vanuit verschillende disciplines een gezamenlijk betekeniskader wordt opgebouwd. Dit is zeker niet de makkelijkste weg, juist omdat er een brug moet worden geslagen tussen verschillende denkwerelden, maar een dergelijke aanpak heeft naar mijn overtuiging wel de grootste wetenschappelijke en maatschappelijke potentie.

7. DANKWOORD

Tot slot wil ik graag enkele woorden van dank uitspreken. Allereerst gaat mijn dank uit naar de leden van het college van bestuur van de Radboud Universiteit Nijmegen voor het in mij gestelde vertrouwen. Verder wil ik het bestuur van de stichting Lorentz van Iterson Fonds TNO en haar raad van advies bedanken voor de verankering van mijn leerstoel in de TNO-organisatie.

Er zijn een aantal mensen die zich in het bijzonder hebben ingespannen voor de realisatie van de leerstoel, vanuit de Radboud Universiteit zijn dat Theo van der Weide en Bart Jacobs, vanuit TNO werd ik vooral bijgestaan door Erik Fledderus en wist ik me gesterkt door de morele ondersteuning van Stephan Raaijmakers en David van Leeuwen. Dank voor jullie beslissende bijdragen. Dank ook aan Arie van Tol, Wilfried Post en Anita Cremers, jullie hebben me op het spoor van een LIFT-leerstoel gezet en zijn dus medeverantwoordelijk voor het feit dat ik hier nu sta.

De kiemen voor mijn carrière als onderzoeker werden gelegd tijdens mijn stage en afstudeerperiode bij het voormalige Instituut voor Perceptie Onderzoek in Eindhoven. Ik ben daarom bijzonder vereerd door de aanwezigheid van professor van Nes. Mijn daaropvolgende aanstellingen bij de letterenfaculteiten van de Tilburgse en Utrechtse universiteit wekten in mij de interesse voor taal en tekst als onderzoeksgebied. Bij TNO kreeg ik altijd de ruimte om mijn academische kant te blijven ontwikkelen. Ik ben TNO zeer erkentelijk voor die vrijheid.

Verder wil ik graag enkele mensen uit het veld bedanken die invloed hebben gehad op mijn ontwikkeling in brede zin doordat we intensief hebben samengewerkt. Allereerst Franciska de Jong, vanaf het begin betrokken bij het IR-onderzoek in Utrecht, jarenlang collega bij TNO, promotor, ik heb veel aan je te danken. Djoerd Hiemstra wil ik in het bijzonder bedanken voor de enerverende samenwerking ten tijde van onze succesvolle TREC-deelnames en de mooie publicaties die daaruit zijn voortgekomen, Arjen de Vries voor de succesvolle samenwerking rond de organisatie van SIGIR 2007, Alan Smeaton, Paul Over en Donna Harman voor de zeer leerzame TREC- en TRECVID-context.

Uiteraard wil ik ook graag mijn ouders bedanken. Jullie hebben het me mogelijk gemaakt om me in de breedte te ontwikkelen. Ik dank jullie voor jullie vertrouwen maar ook voor de nieuwsgierigheid, het geduld en de volharding die ik van jullie heb meegekregen. Ik ben dankbaar dat jullie op deze dag aanwezig kunnen zijn.

Mijn grootste erkentelijkheid gaat uit naar mijn vrouw. Lyne, jouw onvoorwaardelijke steun, vertrouwen en warmte hebben het me mogelijk hebben gemaakt om me voor te kunnen bereiden gedurende de verschillende fases op weg naar dit moment, mijn dank is groot. Ruben en Gaël – onze prachtige kinderen – jullie verwondering is een grote inspiratiebron voor me. Ik ben benieuwd of jullie het later ook leuk zullen vinden om met woorden te spelen.

Ik heb gezegd

NOTEN

1 De Viola browser werd ontwikkeld door Pei-Yuan Wei aan de UC Berkerley. Historische informatie over het project is te vinden op www.viola.org.

2 Usenet.

3 De eerste implementatie is Xanadu, uit 1972. Theodor Holm Nelson. 'As We Will Think'. In: *Proceedings of Online 72 Conference*, Brunel University, Uxbridge, England, 1972.

4 Atkinson, R.D., 'The past and future of America's economy, long waves of innovation that power cycles of growth', *Edward Elgar publishing*, 2004.

5 Lyman, P., Varian, H.R., "How Much Information", 2003 http://www.sims.berkeley.edu/how-much-info-2003 .

6 Vir sapiens fortis est nam et ipsa scientia potestas est. (Francis Bacon, Meditationes Sacrae, 1597).

7 Boyd, D., Ínformation access in a networked world', http://www.danah.org/papers/talks/Pearson2007.html

8 Mul, J. de, 'Overlast en overleven', in *De draagbare lichtheid van het bestaan (Frissen en de Mul red.)* uitgeverij Klement 2008.

9 Wet bewaarplicht telecommunicatiegegegevens http://www.eerstekamer.nl/behandeling/20090828/publicatie_inwerkingtreding/f=y.pdf

10 http://www.ict2030.nl/images/Downloads/ICT2030nl-pres-LR.pdf.

11 *Bush, V., 'As we may think',* The Atlantic Monthly, 1945.

12 Het invloedrijke noisy channel model is gepubliceerd in: Shannon, C.E., "A mathematical theory of communication,'' *Bell System Technical Journal*, vol. 27, pp. 379-423 and 623-656, July and October, 1948

13 Weaver, W., *Translation*, Written 15 July 1949. Published in: *Machine translation of languages: fourteen essays*, ed. by William N. Locke and A. Donald Booth (Technology Press of the Massachusetts Institute of Technology, Cambridge, Mass., and John Wiley & Sons, Inc., New York, 1955), p.15-23.

14 'Bag-of-words' is een gangbaar begrip in de information retrieval en duidt op de vereenvoudigende veronderstelling dat woorden niet van elkaar afhankelijk zijn en woordvolgorde geen betekenis heeft. (Harris, Zellig (1954). "Distributional Structure". *Word* 10 (2/3): 146–62. "And this stock of combinations of elements becomes a factor in the way later choices are made ... for language is not merely a bag of words but a tool with particular properties which have been fashioned in the course of its use".)

15 Belkin, N., 'Some Grand Challenges for Information Retrieval', *ACM SIGIR Forum* 42 (1) 2008.

16 Een gunstige uitzondering is Wikipedia.

17 Zie bijvoorbeeld http://www.w3.org/Consortium/Points/.

18 Gil, Y., Artz, D., 'Towards content trust of web resources', *Proceedings of the 15th international conference on World Wide Web*, 2006.

19 Massa, P., Hayes, C., 'Page-rerank: using trusted links to re-rank authority', *Web Intelligence*, 2005.

20 Brin, S., Page, L., 'Anatomy of a large scale hypertextual Web search engine', *Computer Networks and ISDN Systems*, 30(1-7).

21 Shen, X., Tan, B., Zhai, C., 'Privacy Protection in Personalized Search', *ACM SIGIR Forum*, 41(1).

22 Sassi, F., 'Calculating QALYs, comparing QALY and DALY calculations', *Health and Policy Planning* 2006 21(5).

23 Een voorbeeld van een geparametriseerde kostenmodel is ontwikkeld door het Amerikaanse NIST, Martin, A. et al. "The DET Curve in Assessment of Detection Task Performance", EuroSpeech 1997 Proceedings Volume #4, pp. 1895-1898

24 Hes, R., Borking, J., Privacy Enhancing *Technologies: the path to anonymity (Revised Edition)* Registratiekamer, september 1998. Achtergrondstudies en Verkenningen.

25 Darby, S., *The effectiveness of feedback on energy consumption*, Environmental Change Institute, University of Oxford, 2006.

26 http://www.nrc.nl/binnenland/article2180267.ece/Arrestatie_na_dreiging_met_schietpartij_Breda