Van telecommunicatie naar semantisch web

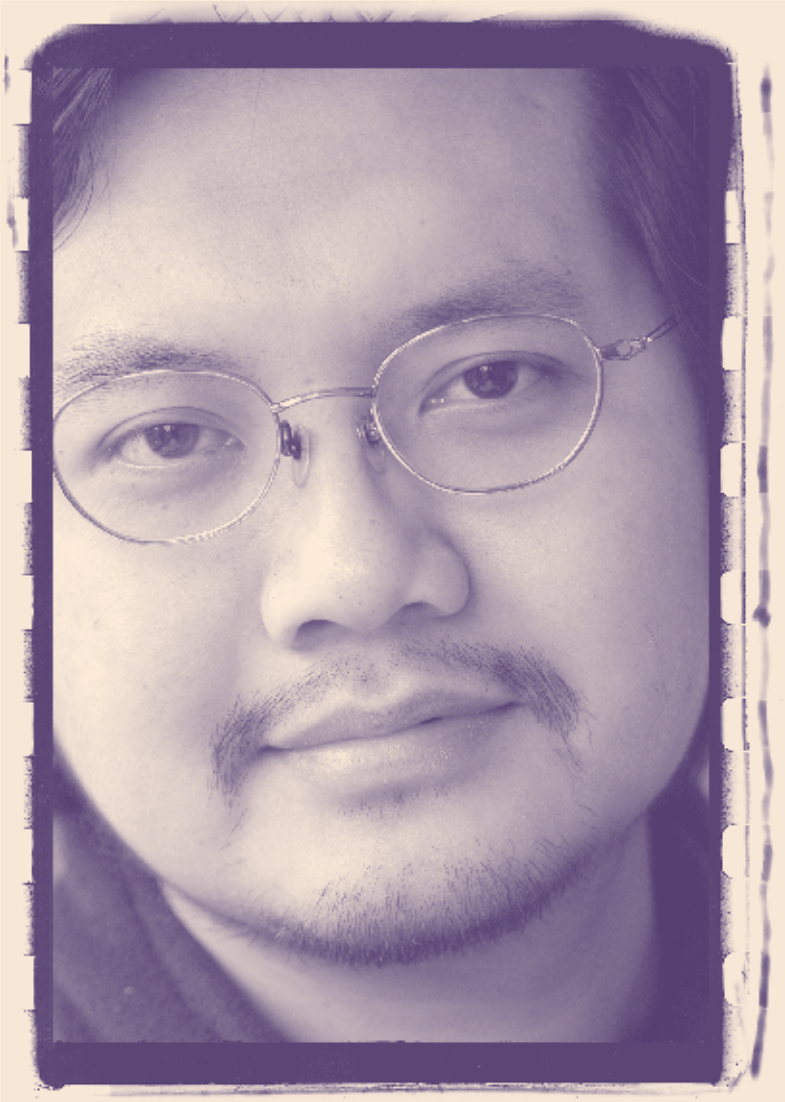
het interview

Frank den Hollander f.j.den.hollander@rug.nl
Kristien Piersma k.i.piersma@rug.nl

Met dank aan Peter van Laarhoven

Fotografie: Gerhard Lugard

Ismail Fahmi (32) heeft een bijzondere positie: hij werkt als AIO bij Alfa-Informatica, maar is ook voor de helft in dienst van de Universiteitsbibliotheek. Een gesprek over ontologieën, het semantisch web en Indonesië.



Ismail Fahmi

Ik ben geboren in Bojonegoro op Oost-Java. Ik ging studeren op West-Java, in Bandung, dat is achthonderd kilometer verderop! Ik heb aan de universiteit van Bandung de bacheloropleiding telecommunicatie gedaan, dus met antennes en frequenties en zo. In 1997 was ik klaar en ben ik gaan werken aan de digitale bibliotheek. In die tijd was die er nog niet in Indonesië, dus ben ik er een gaan opzetten in Bandung, alsmede een community en een netwerk eromheen.

In het kader van het samenwerkingsverband tussen de RUG en de universiteit van Bandung waren bibliothecaris Alex Klugkist en Madeleine Gardeur van het Bureau Buitenland van de RUG in 2000 in Indonesië. Toen hebben we gepraat over een beurs voor mij in Groningen, en Alex Klugkist heeft toen geprobeerd een hoogleraar hier te vinden waar ik bij kon promoveren. Dat werd John Nerbonne, en nu heb ik een speciale aanstelling: voor vijftig procent werk ik bij de Universiteitsbibliotheek, en de andere vijftig procent besteed ik aan mijn onderzoek. Normaal wordt een AIO betaald uit een project in de faculteit, maar ik word dus betaald door de UB.

Eerst werkte ik met Peter van Laarhoven aan een project over de gebruiksstatistieken van elektronische tijdschriften. Dat was

niet direct gerelateerd aan mijn onderzoek, toen had ik dus twee verschillende aandachtsvelden. Nu werk ik binnen de nieuwe afdeling Digitale Bibliotheek onder leiding van Henk Ellermann, waardoor de ideale situatie is ontstaan dat ik ook hier werk doe dat direct relevant is voor mijn onderzoek bij de vakgroep.

Term extraction

Bij Alfa-Informatica houd ik me bezig met het extraheren van belangrijke termen en terminologie uit documenten. Stel, je bekijkt het artikel over *PsychiatryOnline* dat in deze Pictogram staat (pag. 15, red.): de computer is niet in staat om de belangrijkste termen daaruit te destilleren. Ik werk met een corpus van medische documenten en probeer het systeem in staat te stellen medische termen uit die documenten te trekken, die dat document goed beschrijven. Het is niet hetzelfde als metadata, dat is wat van buitenaf gedefinieerd is: de titel, de auteur, trefwoorden etc. Het gaat hier om termen ín de tekst.

Er zijn verschillende technieken, die gebruik maken van wiskundige statistische methoden, zoals *log-likelihood*, en methodes die speciaal ontwikkeld zijn voor *term extraction*. Een term als *APA Practice Guidelines*, uit het artikel over PsychiatryOnline, bestaat uit drie woorden die samen een tekst-string vormen. Diezelfde tekst-string komt waarschijnlijk ook nog ergens anders in de tekst voor. Als we nu deze *keywords* op verschillende plaatsen kunnen traceren, kunnen we relaties leggen met de context eromheen.

We gebruiken de wiskundige methoden om de frequentie te tellen. Een hoge frequentie zegt niet alles, want je kunt bijvoorbeeld verwachten dat een tekst-string als 'het is' in Nederlandse teksten erg veel voorkomt. Onze methode ziet het verschil tussen zo'n 'stopwoord' en echte termen die vaak voorkomen.

Hoe weet je of een term relevant is?

Je kunt een lijst van standaard medische termen gebruiken, maar dan is er een probleem als er een nieuwe term bijkomt, voor een nieuw ziektebeeld of nieuw medicijn bijvoorbeeld. We proberen alleen met statistische methodes te kijken naar veelvoorkomende termen binnen een bepaald domein. 'Afasie' zal meer in medische documenten voorkomen en minder vaak in een historisch domein. Dus doen we vooral onderzoek binnen één domein.

Als we een corpus met teksten hebben, is het niet reëel dat in elk document het woord afasie voorkomt. Misschien zit het slechts in tien documenten. Als we nu het woord meerdere malen in een bepaald document vinden, ligt het voor de hand dat afasie een term is die bij dit document hoort, die het beschrijft.

Waarvoor wordt deze technologie gebruikt?

Met de enorme groei van het onderzoek en de toename van het aantal dissertaties krijg je ook steeds meer nieuwe medische technieken en dus ook meer medische terminologie. Elk nieuw medicijn, elke nieuwe behandeling krijgt een nieuwe naam.

Als we alleen van menskracht afhankelijk zijn, moeten al die documenten, artikelen, dissertaties allemaal gelezen worden om nieuwe termen te ontdekken. Dat is niet praktisch. Maar als een computer een document kan 'lezen' en er alle belangrijke terminologie uit kan destilleren, scheelt dat veel zoekwerk.

Een andere toepassing is het gebruik bij *information retrieval*. Als mensen bijvoorbeeld via Google naar documenten zoeken, tikken ze de belangrijkste zoektermen in. Maar ons systeem is in staat termen in documenten te vinden en te zeggen: deze termen beschrijven dit document het beste. Als een gebruiker dan met die term gaat zoeken, krijgt het dus die documenten eerst die die term het best beschrijven, die het meest relevant zijn.

Bij Alfa-Informatica produceer ik alleen maar een methode en theorie, en houd ik me niet bezig met een toepassing ervan. Maar in de bibliotheek moet ik een applicatie afleveren die gebruikt kan worden door de bezoekers. Hun kan het niet schelen wat voor theorie of methode erachter zit; het moet nuttig zijn!

Semantisch web

Het doel van mijn werk in de bibliotheek is een stuk van het semantisch web te ontwikkelen, en wel speciaal voor een historisch domein. Om uit te leggen wat het semantisch web is, moeten we even in de historie van het internet duiken.

We zijn inmiddels aan de derde generatie van het internet toe. In de eerste generatie, het begin van het internet, had je alleen statische pagina's, zonder scripts en dergelijke. De tweede generatie is die van de databases achter veel websites: er ontstonden dynamische pagina's. Toch was het formaat, het uiterlijk van een webpagina, voornamelijk bedoeld voor directe consumptie door gebruikers. We kunnen de titel lezen,

> Nederland is erg leuk, ik heb totaal geen heimwee! <

de illustraties bekijken, et cetera. Maar computers begrijpen niets van zo'n site! Hooguit kan een computer webpagina's indexeren, proberen keywords te vinden en algoritmes toe te passen om waarden toe te kennen, zoals Google nu doet, voor information retrieval.

Maar vraag je Google: *waarom* werd president Lincoln vermoord? Dan zal Google niet kunnen antwoorden, tenzij je precies de goede zoektermen hebt ingetikt en er een pagina opduikt die precies de goede relevantie heeft.

In de derde generatie van het internet, die we dus het semantisch web noemen, gaan we ervan uit dat er een dienstverlening ontstaat waarbij je om specifieke informatie kunt vragen: 'Wat was de betekenis van "democratie" in de zeventiende eeuw?' en het systeem zal die informatie kunnen geven met behulp van zogenaamde ontologieën –waarbij het gaat om het definiëren van de goede termen en hun context.

Democratie in de 17e eeuw

Een voorbeeld. Een van de bronnen die we gebruiken is de *Evans Early American Imprints*. Dat is een grote verzameling met alle vroege Amerikaanse drukken gepubliceerd tussen 1600 en 1800: van brieven, liederen, speeches en pamfletten tot hele boeken. Deze is door Letteren en de UB gekocht voor de studie Amerikanistiek en wordt gewoon via het web geraadpleegd bij de uitgever. Wat wij doen is werken met de gescande teksten in ASCII-formaat, die door de UB zijn gekocht als archiefversie; dat is een enorm corpus, dat zich goed leent voor omzetting naar onze werkformaten.

De Amerikanisten aan deze universiteit zijn geïnteresseerd in dit materiaal en ze willen weten hoe ons systeem hen kan helpen de geschiedenis beter te onderzoeken. Het semantisch web nu stelt ons in staat niet alleen de informatie in de metadata te onderzoeken, maar ook de informatie in het document zelf. In een document staat bijvoorbeeld iets over een bepaald persoon, en die doet iets met een andere persoon: ze hebben een probleem, ze communiceren met elkaar of vermoorden elkaar. Deze informatie wordt nu gecodeerd in een stramien, en dat noemen we een ontologie. In een ontologie zit een structuur, bijvoorbeeld "persoon - locatie - geschiedenis - organisatie - gebeurtenis".

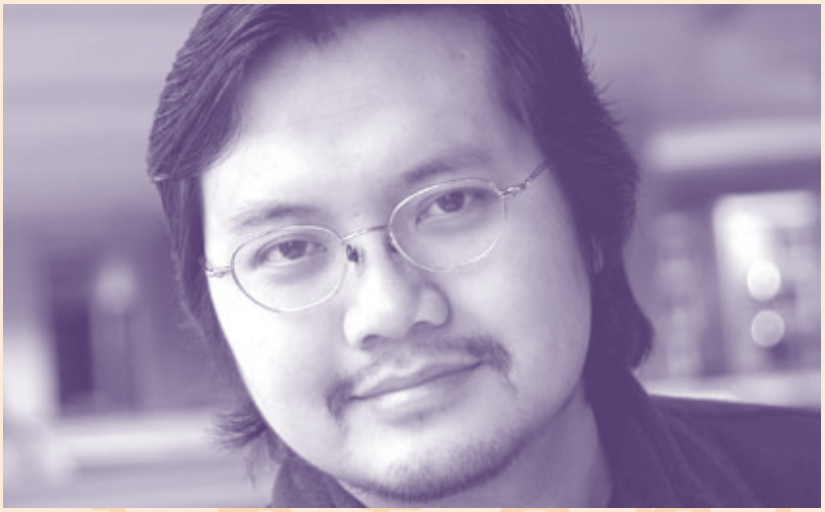
Een historicus wil bijvoorbeeld iets weten over het gebruik van het woord democratie. Is er een verschil van betekenis tussen het woord democratie dat nu wordt gebruikt, en het woord zoals het werd gebruikt in de 17e eeuw? Het systeem zou nu het woord democratie moeten kunnen extraheren, en ook de context kunnen duiden in een specifiek jaar. De ene keer wordt het bijvoorbeeld gebruikt door Roosevelt in een speech bij een begrafenis, en in een ander jaar bij een rechtszaak. En misschien wordt het een paar jaar later weer in een andere context gebruikt. Op deze manier kunnen mensen betekenisverschuivingen van een woord analyseren in een bepaalde periode. In een oogopslag kan men op een tijdbalk het gebruik van zo'n woord volgen. Maar we kunnen ook zoeken op gebeurtenis, en dan in bepaalde context.

Definition extraction

Mijn volgende taak hier zal zijn om definities bij de termen te vinden. Er zijn natuurlijk al meer mensen die ditzelfde onderzoek doen; er is de ACL, de *Association for Computational Linguistics*. Wat ik nu doe, *definition extraction*, is het nieuwste op dit gebied.

De techniek is er al, de tools zijn er, maar mijn taak is om ze toe te

passen in de bibliotheek, te beginnen bij de Evans database. Als we in de toekomst duidelijk hebben hoe het semantisch web werkt en er meer ervaringen mee hebben, gaan we het uitbreiden naar andere bronnen. Want de kracht van het semantisch web is om alle informatie die voorhanden is te combineren. Een gebruiker wil bijvoorbeeld het woord democratie vinden in die vroege Amerikaanse drukken. Het semantisch web heeft aan slechts één bron genoeg om feedback te geven, een tijdbalk etc. Maar we hebben ook de boekencatalogus en de repositories, databases die in de toekomst deel kunnen gaan uitmaken van het semantisch web. We kunnen informatie onttrekken aan welk documenttype dan ook, en het opnemen in een ontologie.



Dit alles is nog in de proefondervindelijke fase, we kunnen nu op een zeer ruwe manier zo'n soort ontologie presenteren, maar het vergt nog veel research en ontwikkeling om het te perfectioneren. Als we dat allemaal voor elkaar hebben, en je komt weer met het trefwoord democratie, dan zal het systeem niet alleen met die vroege Amerikaanse bronnen komen, maar ook met andere documenten en vooral: andere onderzoekers die met hetzelfde onderwerp bezig zijn! Referenties, citaten, verwijzingen, links, ga zo maar door.
De uitdaging is dat er nog niet veel toepassingen zijn voor het semantisch web. Er is veel achtergrondkennis nodig over ontologie, *formatting* en dergelijke. Het project is nog maar net begonnen. We hopen over ongeveer een jaar een eerste versie van het semantisch web klaar te hebben, waarna we nog een jaar nodig hebben om te testen en evalueren.

Bevalt het je in Nederland?
Dat ik in Groningen terechtkwam, was mede te danken aan het bezoek van Alex Klugkist aan Bandung. Maar ik heb hier ook veel andere contacten, er zijn veel Indonesische studenten hier. Ik kon ook een beurs krijgen voor Japan, maar daar wilde ik niet heen. Er zijn hier meer landen en culturen, het is hier meer menselijk in Europa. Ik ben tien dagen in Japan geweest, maar was blij dat ik weer terug was in Indonesië.
Nederland vind ik erg leuk, ik heb totaal geen heimwee! Mijn vrouw en twee kinderen zijn hier ook, en die vinden het ook geweldig hier. De kinderen spreken al erg goed Nederlands, ze leren ons nieuwe woorden. Mijn vrouw en ik zijn nu op niveau 2 met Nederlandse les, bij het Talencentrum.

Tsunami
We hebben een heel actieve studentenorganisatie in Groningen, ik ben er in twee erg actief: PPI, de Indonesische studentenorganisatie, en Degromiest, een gemeenschap voor Indonesische moslims in Groningen, daar was ik nog een tijdje voorzitter van.
Na de tsunami die met kerst 2004 onder andere Indonesië trof, konden we makkelijk mensen mobiliseren hier in Groningen en toen hebben we de website *Aceh Update* opgezet. We werkten samen met een team journalisten in Indonesië, die informatie verzamelden over de tsunami, van het begin tot enkele dagen na de vloedgolf. Ze probeerden deze nieuwsberichten in zoveel mogelijk talen de wereld over te sturen. Wij konden gratis via RSS-feeds hun informatie op onze website zetten. Het gaat voornamelijk om journalistieke items, die vertellen over de stand van zaken, verhalen van slachtoffers en dergelijke.
Ik heb geen enkel probleem als moslim in Nederland. Ik hoor wel eens van mensen dat ze bang zijn, het klimaat is toch wel harder geworden. Je hoort over nieuwe regels, dat men niet meer in een burka over straat mag gaan... Maar ik kan hier op de UB bijvoorbeeld gewoon bidden wanneer ik dat wil, in het vergaderkamertje vlakbij mijn werkplek, en als dat bezet is vind ik wel ergens anders een plekje. Misschien is het voor moslimstudenten hier in de bibliotheek wel moeilijk... maar aan de andere kant: er zijn genoeg plekjes tussen de boekenkasten!
Ik wil nog wel graag iets kwijt over mijn persoonlijke gevoel over werken en leven aan deze universiteit: ik vind het erg fijn om hier te kunnen studeren, ik hou van de omgeving en de mensen, en ik prijs me gelukkig dat ik van zoveel mensen hulp krijg! Dat maakt dat ik me zowel op het werk als erbuiten erg thuis voel.

Links:

- Het project over gebruiksstatistieken van elektronische tijdschriften waaraan Ismail Fahmi meewerkte, resulteerde in een presentatie op de LIBER conferentie: *Usage statistics of online journals: background, trends & prospects - with a local elaboration.*
  De bijbehorende Powerpointpresentatie:
  http://liber.ub.rug.nl/presentations/Laarhoven.ppt
- De door Ismail Fahmi beheerde website van de gemeenschap van Indonesische moslims in Groningen, Degromiest:
  http://cafe.degromiest.nl
- Ismails website over de tsunami:
  http://acehupdate.degromiest.nl