Bernd Wondergem, Patrick van Bommel, Theo Huibers en Theo van der Weide

De elektronische informatiemakelaar

Vernieuwd! Verbeterd! Nu met extra witbeschermer, of zelfs gebaseerd op liposomen. Kreten over wasmiddelen en schoonheidscrèmepjes die aangeven dat het product ingrijpend verbeterd is. Of toch slechts simpelweg reclamepraat? Iets gelijks lijkt er aan de hand in de wereld van zoeksystemen en Information Retrieval. Agents! Proactief! Intelligent! Zijn het loze kreten die alleen goed klinken en goed verkopen? Trouwens, wat zijn agenten eigenlijk. Hoe ziet de opbouw van een zoeksysteem gebaseerd op agenten eruit. Leidt het gebruik van agenten echt tot verbetering van zoeksystemen? Een checklist van eigenschappen voor het begrip 'agent'.

Information Retrieval (1) krijgt zo wel in het bedrijfsleven (Documentaire Informatie Systemen, full-text retrieval enzovoort) als in de universitaire onder-zoekswereld grote aandacht. Information Retrieval (IR) is steeds belangrijker geworden, met name door de groeiende populariteit van het Internet. Dit blijkt bijvoorbeeld uit de opkomst van vele zoekmachines, zoals AltaVista, Lycos, Infoseek en de speciaal voor de Nederlandse markt gemaakte search engines Ilse en Vindex. Bestaande IR-systemen zijn echter verre van ideaal, waardoor verder onderzoek met zowel wetenschappelijke als commerciële doelen noodzakelijk is.

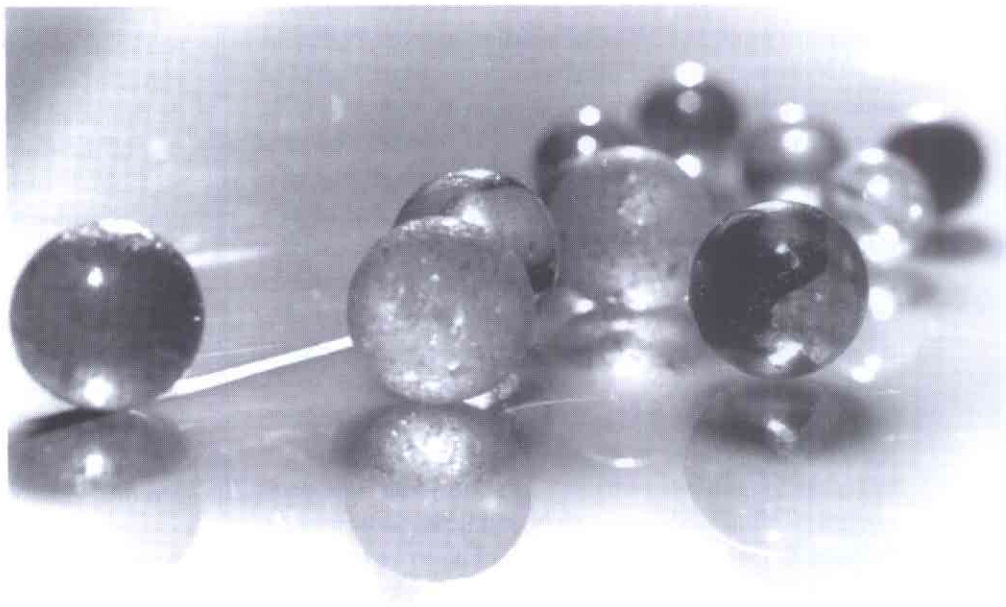


Foto: Egon Viebre

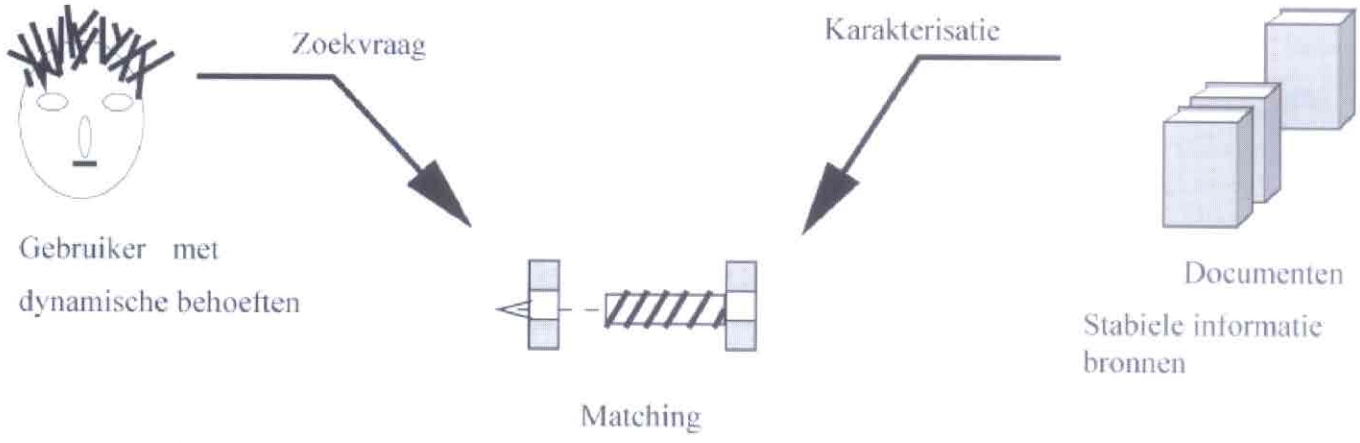
De huidige tekortkomingen van IR-systemen kunnen worden onderverdeeld in drie categorieën. Dit gebeurt aan de hand van de drie hoofdtaken van een IR-systeem (zie figuur 1): (1) interactie met en modellering van gebruikers, (2) indexering van documenten en (3) matching.

Gebruikers

Ten eerste zijn er problemen die direct met gebruikers te maken hebben. Het formuleren van een zoekvraag is nog steeds geen sinecure. Bij deze taak zouden IR-systemen hun ondersteunende rol beter moeten vervullen. *Queries* kunnen met verschillende doelen gesteld worden, bijvoorbeeld fact-finding, explorerend, verzamelend. Deze verschillende modaliteiten zouden door het IR-systeem onderscheiden moeten worden.

Dit hangt ook samen met de manier waarop de gevonden documenten aan de gebruiker gepresenteerd worden. Het IR-systeem zou de voorkeuren van de gebruiker hieromtrent moeten achterhalen. Op basis daarvan kan besloten worden de documenten te presenteren aan de hand van, bijvoorbeeld, een ranked list, clusters of kennisgraaf. Ook ergonomische aspecten zoals het ge-

Drs. Bernd Wondergem, dr. Patrick van Bommel en dr.ir. Theo van der Weide zijn verbonden aan het Computing Science Institute van de Katholieke Universiteit Nijmegen. Dr. Theo Huibers is werkzaam bij DOXiS Documentaire Informatiespecialisten te Leidschendam.



Figuur 1. Information Retrieval paradigma.

bruik van kleuren en de indeling van het scherm zouden moeten worden toegesneden op de individuele gebruiker. Bij de interactie met gebruikers moet bovendien de privacy gewaarborgd worden.
Kortom, aan de gebruikerskant kan met personalisatie veel verbeterd worden. Dit betekent dat het zoeksysteem zich aanpast aan de specifieke wensen van de individuele gebruiker.

Documenten

Ten tweede zijn er problemen die met documenten te maken hebben. Vergeleken met de traditionele IR-context, zoals bijvoorbeeld een bibliotheeksysteem, is de dynamiek tegenwoordig veel hoger. Het vormt een groot probleem dat er op elk moment documenten verdwijnen, worden aangepast, of geheel nieuw beschikbaar worden gesteld.
Bij de indexering van documenten kunnen verschillende technieken gebruikt worden. Naast de gebruikelijke keywords bestaan meer geavanceerde descriptoren als index-expressies en noun-phrases.

Matching

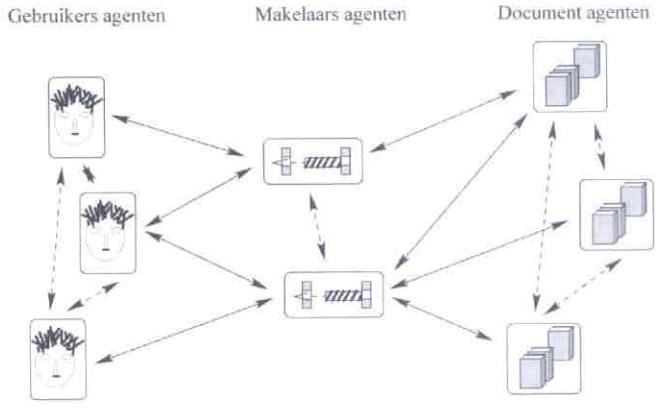
Tot slot zijn er problemen die te maken hebben met de koppeling tussen gebruikers en documenten, i.e., matching. Ook hier klinkt de roep om een gepersonaliseerde aanpak.
Een vorm van personalisatie is adaptieve matching. Adaptieve matching kan verkregen worden door het gebruik van lerende algoritmen. Om de effectiviteit van lerende algoritmen naar een hoger plan te tillen is het nodig dat de gebruiker inzicht in het proces van relevantiebepaling heeft. De mogelijkheid het IR-systeem gerichte en gedetailleerde informatie te verschaffen, stelt de gebruiker dan in staat de werking van het IR-systeem beter naar zijn hand te zetten. Een typische manier om dit inzicht te verschaffen is in een dialoog met het systeem.

Informatiemakelaar

De genoemde problemen zijn aan te pakken met zogenaamde agenten. Voor de drie hoofdtaken van IR-systemen worden verschillende agenten gevormd: gebruikersagenten, documentagenten en makelaarsagenten. Tussen deze agenten kan communicatie plaatsvinden, zoals geschetst in figuur 2.
Gebruikersagenten vormen een uitgebreid beeld van de interessen van de gebruiker, een gebruikersprofiel. Naast de inhoudelijke beschrijving van de interessegebieden bevat het profiel ook informatie over, bijvoorbeeld, de talen die de gebruiker spreekt (intypt als zoekvraag of kan lezen).
Documentagenten bekijken (bronnen met) documenten en maken een overzicht van het (veranderende) aanbod. Door middel van indexering vormen zij beschrijvingen van de inhoud van documenten. Ook hier kunnen additionele kenmerken over bijvoorbeeld grootte en plaats meegenomen worden.
Tussen de gebruikersagenten en documentagenten wordt de derde soort agent geplaatst: de makelaarsagent. Makelaarsagenten vervullen de functie van informatiemakelaar. Ze werken als tussenpersoon en zorgen voor de informatieoverdracht tussen gebruikers en documenten. De belangrijkste functie van makelaarsagenten is het bepalen welke documenten relevant zijn voor de gebruikers en vice versa. Het gebruik van dergelijke agenten als tussenpersoon levert ook de basis om het privacyprobleem op te lossen. Een zoekvraag van de gebruiker wordt behandeld door de tussenpersoon die de documentbron bevraagt zonder de gegevens (naam, adres enzovoort) van de gebruiker door te spelen.

Agents

Er zijn verschillende definities of beschrijvingen van agenten in omloop. Deze variëren van simpele zoals 'een entiteit die doelen nastreeft', tot vrij uitgebreide als de onderstaande.



Figuur 2. Multi-agent-systeem voor IR.

Uit het grote aanbod van definities zal een bruikbare gekozen moeten worden. Een bruikbare definitie is zodanig dat belangrijke elementen uit IR ermee beschreven (kunnen) worden. De omschrijving van Wooldridge & Jennings (4) is zeer bruikbaar voor IR. Deze omschrijving noemt de volgende vijf eigenschappen als essentieel voor agenten: autonoom, reactief, proactief, sociaal en intelligent. We zullen deze eigenschappen beschrijven in termen van IR en aangeven hoe ze de geschetste problemen (deels) kunnen oplossen.

Autonoom

Autonomie wil zeggen dat agenten een zekere vrijheid hebben in hun doen en laten. De agent kan, tot op zekere hoogte, zelf bepalen welke acties hij wanneer zal uitvoeren. Een makelaarsagent kan beslissen eerst documentagenten te consulteren om meer informatie over bronnen te verkrijgen alvorens zoekvragen van gebruikersagenten te behandelen.

Reactief

Reageren op prikkels uit de omgeving wordt reactiviteit genoemd. Tot de omgeving van agenten rekenen we andere agenten en eindgebruikers. Praktisch alle informatiesystemen zijn reactief omdat ze reageren op verzoeken van gebruikers.

Proactief

Proactiviteit stelt dat een agent zonder directe opdracht het initiatief kan nemen. Gebruikersagenten proberen repeterende zoekopdrachten te herkennen. Denk bijvoorbeeld aan iemand die vlak voordat hij van zijn werk met de auto naar huis gaat even op het Internet kijkt waar de files staan. Een proactieve gebruikersagent zal, na herkenning van dit patroon, niet wachten tot de opdracht weer gegeven wordt. Hij zal zorgen dat er tijdig een bericht (bijvoorbeeld middels e-mail) naar de gebruiker wordt gestuurd met daarin een recent fileoverzicht.

Proactiviteit onderscheidt agenten van veel bestaande zoeksystemen. Sommige zoeksystemen tonen een beperkte variant van proactiviteit. Bijvoorbeeld de Informant: 'Your personal search agent on the Internet'.
De proactiviteit van de Informant gaat niet verder dan de mogelijkheid tot een periodieke herhaling van een zoekvraag. De Informant zal dus niet spontaan het initiatief nemen, maar slechts proactief zijn volgens een door de gebruiker aangegeven interval. Deze intervalproactiviteit is op dit moment de state-of-the-art.

Sociaal

Agenten zijn sociaal omdat ze kunnen communiceren met hun omgeving en met name met andere agenten daarin. Deze eigenschap is essentieel in multi-agentsystemen. Communicatie stelt agenten in staat samen te werken (3).
In het multi-agent-systeem voor IR van figuur 2 is communicatie onontbeerlijk. De drie typen agenten vormen samen het informatiesysteem waarbinnen communicatie nodig is om relevante documenten uit bronnen via makelaars bij gebruikers af te leveren.
Zoals tevens in figuur 2 te zien is, kunnen agenten van hetzelfde type ook onderling communiceren. Dit stelt makelaarsagenten in staat om relevantiebepalingen van andere makelaars mee te nemen. Voorbeelden hiervan zijn meta-zoeksystemen als MetaSearch (www.metasearch.com) en MetaCrawler (www.metacrawler.com). Onder het motto 'twee weten meer dan één' verspreiden zij de ingekomen zoekvraag onder een aantal (standaard-)zoeksystemen en voegen de resultaten hiervan op een bepaalde wijze samen.

Intelligent

Begrippen die sterk met intelligentie te maken hebben, zijn leren en adaptatie. Intelligentie bij agenten kan verkregen worden door toepassing van geavanceerde technieken uit de kunstmatige intelligentie. Belangrijke technieken voor IR zijn natuurlijke taal verwerking (NLP) en lerende algoritmen voor adaptieve matching. Autonomy's software (www.autonomy.com/) stelt de gebruiker in staat expliciete relevance feedback te geven. Dit gebeurt met behulp van de zogenaamde retrainfunctie. Op grond hiervan worden relevantiebepalingen bijgesteld en aangepast aan de wensen van de gebruiker.

Profile

Het Profile-project (2) is een onderzoeksproject aan de Katholieke Universiteit Nijmegen. Doel van het Profileproject is het maken van een proactief informatiefilter. Dit gebeurt op basis van het paradigma uit figuur 2.
Samenwerking tussen het Nijmegen Institute of Cognition and Information (NICI) en het Computing Science Institute (CSI) zorgt voor een mix van kennis

over ergonomie, gebruikersmodellering, natuurlijke taalverwerking en vergelijkingsmaten.
Het informatiefilter laat relevante informatie door en filtert irrelevante informatie. Het filter maakt gebruik van gebruikersprofielen om vast te stellen of binnenkomende informatie al dan niet relevant is voor de gebruiker. Het profiel van de gebruiker bevat naast een omvangrijke beschrijving van zijn interessegebieden een aantal situationele factoren. Deze beschrijven additionele kenmerken van de gebruiker zoals leeftijd, beroep en de talen die hij beheerst.
Het gebruikersprofiel wordt proactief geconstrueerd. Dit houdt in dat het Profile-systeem 'over de schouder' van de gebruiker meekijkt en uit diens acties het profiel afleidt. Bij de gebruikersmodellering wordt gebruikgemaakt van ontologieën die bepaalde domeinen beschrijven.
Binnenkomende documenten worden door een *parser* ontleed. Dit levert een beschrijving van de inhoud van het document in de vorm van een aantal descriptoren. Binnen het Profile-project worden verschillende smaken descriptoren onderzocht. Naast de veelgebruikte termen (keywords) worden ook index-expressies en noun-phrases gebruikt. Deze samengestelde descriptoren hebben een grotere uitdrukkingskracht en bevatten (taalkundige) structuur. Normalisaties worden op deze specifieke descriptoren toegepast om semantisch equivalente expressies te identificeren.
Nieuwe vergelijkingstechnieken worden binnen Profile ontwikkeld om de mate van overeenkomst tussen descriptoren vast te stellen. Momenteel wordt hard gewerkt aan een vergelijkingsmaat voor zogenaamde Booleaanse index-expressies. Deze combineren structuur die de taalkundige semantiek weergeeft met logische operatoren en redeneerkracht.
Het Profile-project heeft inmiddels geresulteerd in een aantal artikelen. Tevens is een eerste prototype geïmplementeerd in Java. In het prototype communiceren de drie genoemde typen agenten via een router. Een volgende versie zal verder uitgewerkte agenten bevatten. Naast software voor het beoogde systeem worden er ook programma's geschreven om aspecten te testen. Voorbeelden hiervan zijn een testomgeving voor lerende classificatie-algoritmen en een implementatie van Booleaanse index-expressies in de functionele taal Clean.
Het vierjarige project loopt tot en met september 2000. Informatie over het Profile-project kan op het Internet gevonden worden op de Profile-homepage http://hwr.nici.kun.nl/~profile/. Gerelateerde artikelen zijn beschikbaar op www.cs.kun.nl/~bernd.

Tot slot

In dit verhaal is behandeld hoe agenten ingezet kunnen worden voor Information Retrieval. Op basis van de drie hoofdaspecten van IR werden drie typen agenten benoemd: gebruikers-, document- en matchingsagenten. Het laatste type werd verder uitgewerkt in de vorm van informatiemakelaar. Tevens werden vijf eigenschappen besproken die wij kenmerkend vinden voor agenten: autonoom, reactief, proactief, sociaal en intelligent. Tot slot belichtten we het Profile-project.
Het woord agent is tegenwoordig ietwat trendy en het is niet altijd duidelijk wat er precies mee bedoeld wordt. De vijf genoemde eigenschappen zijn nuttig en bruikbaar maar geven een erg brede definitie van het begrip agent. Zij zullen voor toepassing in IR concreter ingevuld moeten worden. Er kunnen bijvoorbeeld verschillende categorieën of soorten proactiviteit onderscheiden worden. Ook een bruikbare omschrijving van intelligentie ontbreekt nog.
Als een concrete checklist beschikbaar is kunnen systemen pas echt op hun agent-merites beoordeeld worden. Dan zal ook blijken of de term agent gebruikt wordt als reclamepraat of als zinnige aanduiding.

The Informant
De Informant (http://informant.dartmouth.edu/) bewaart drie persistente zoekvragen en voert ze periodiek uit. De resultaten hiervan worden middels e-mail aangekondigd. Via de webinterface kunnen de resultaten bekeken worden. De Informant is een gratis service ontwikkeld door Dartmouth College. Bij aanmelding ontvangt men een password dat toegang geeft tot de persoonlijke pagina met resultaten. Men kan drie Booleaanse zoekvragen ingeven en een zoekmachine waarmee de zoekvraag uitgevoerd moet worden. Naast de zoekvragen kunnen ook een aantal URL's opgegeven worden die de Informant in de gaten moet houden. Na wijziging van een van de pagina's wordt men gealarmeerd.

Bibliografie

1. C.J. van Rijsbergen. *Information Retrieval*. Butterworths, London, United Kingdom, 1990.
2. B.C.M. Wondergem, P. van Bommel, T.W.C. Huibers and Th.P. van der Weide. 'Opportunities for Electronic Commerce in Information Discovery'. In: F. Griffel, T. Tu and W. Lamersdorf (eds.), *Proceedings of the International IFIP/GI Working Conference on Trends in Distributed Systems for Electronic Commerce, TrEC 98*, pages 126-136, Hamburg, Germany, June 1998.
3. B.C.M. Wondergem, P. van Bommel and Th.P. van der Weide. 'Cumulative Duality in Designing Information Brokers'. In: *Proceedings of the 9th International Conference on Database and Expert Systems Applications (DEXA)*, Vienna, Austria, August 1998.
4. M. Wooldridge and N.R. Jennings. 'Intelligent Agents: Theory and Practice'. In: *Knowledge Engineering Review*, 10(2):115-152, 1995.