

**Persoonlijk Nieuws: een kijkje in de keuken van de
personalisatie van online nieuwswebsites**
Bachelorscriptie informatiekunde

Koen Hulsman
0616265
Radboud Universiteit Nijmegen

22 juni 2009

Inhoudsopgave

1	Inleiding	4
2	Gepersonaliseerd Nieuws in de Literatuur	5
2.1	Profile Generation	6
2.2	Profile Maintenance	7
2.3	Profile exploitation	9
2.3.1	SEAN	10
2.3.2	Ontologieën	14
2.4	Samenvattend	16
3	Case Studie	18
3.1	Uitleg	18
3.1.1	GoogleNews	19
3.1.2	Spotback News	21
3.1.3	Overeenkomsten	22
3.2	Case Studie Beschrijving	22
3.3	Resultaten	23
3.3.1	Google News	25
3.3.2	Spotback News	27
3.3.3	Winnaar?	28
4	Conclusie	29

1 Inleiding

Al sinds eind jaren '90 is het mogelijk om via het internet gebruikt te maken van persoonlijke nieuwswebsites: websites die jou het nieuws tonen dat jij interessant vindt. De service is onderdeel van de veel grotere stroming van de 'personal services' in de digitale wereld. Dit uit zich niet alleen op het internet, maar ook op mobiele telefoons en pda's [MGB⁺04]. Op internet zijn 'personal services' vaak gekoppeld aan webservices [ACT01]. Webservices zijn services die via het internet toegankelijk zijn en op een server worden uitgevoerd. Enkele voorbeelden van webservices, behalve gepersonaliseerde nieuwssites, zijn websites om vliegtickets te zoeken en boeken of nog simpeler, euro's om te rekenen naar dollars. Het idee achter de 'personal services' is simpel: je past een website aan de voorkeuren van je klant aan. Zo zal een webwinkel bij iemand die van fantasyboeken houdt, proberen zoveel mogelijk fantasyboek gerelateerde items te tonen op de pagina. Achter dit simpele idee schuilt een veel moeilijker probleem: hoe kom je erachter dat iemand van fantasyboeken houdt? Hoe weet je of iemand geïnteresseerd is in nieuwsberichten over bijvoorbeeld binnenlandse politiek?

Dezelfde vragen komen bovendien bij de sites die gepersonaliseerd nieuws aanbieden. Zij proberen namelijk, via directe en indirecte interactie met de gebruiker, zijn/haar voorkeuren te leren kennen, een profiel op te stellen en via dat profiel voor de gebruiker interessante nieuwsberichten te plaatsen.

Vanaf eind jaren '90 zijn er vele sites boven komen drijven op het web. En, vaak ook erg snel weer gezonken. Een van de eerste echt succesvolle gepersonaliseerde nieuwsaanbieders was Yahoo met haar Yahoo News. Yahoo bood deze service aan de gebruikers van hun mail service. Andere sites die op dezelfde manier deze service aanbieden zijn Google en Netscape (AOL). Daarnaast is er de Spotback, een site die gebruikers zelf links naar nieuwsberichten laat plaatsen en deze nieuwsberichten laat raten. Deze ratings worden ook meegenomen als interessante nieuwsberichten voor een gebruiker worden gefilterd. In mijn scriptie zal ik onderzoeken welke methoden worden gebruikt bij het mogelijk maken van gepersonaliseerd nieuws. Ik zal daarvoor eerst kijken welke methoden er in de literatuur te vinden zijn. Daarna zal ik met behulp van deze beschrijving een praktijkstudie uitvoeren op meerdere gepersonaliseerde nieuwssites.

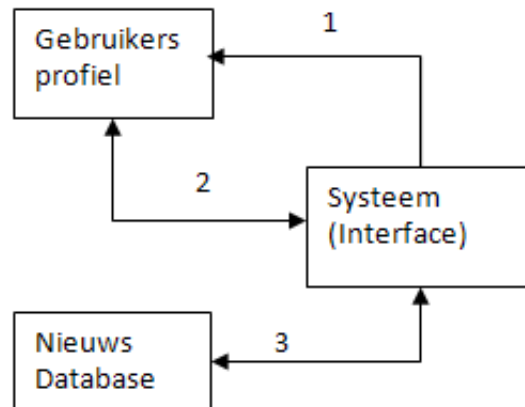
De vraag die centraal zal staan in dit onderzoek is de volgende:

Welke methoden worden er in de literatuur genoemd om voor gepersonaliseerde nieuwssites de voorkeuren van gebruikers te leren kennen, van welke van deze methoden maken vooraanstaande gepersonaliseerde nieuwssites gebruik en hoe goed werken deze sites in de praktijk?

Kijkend naar gepubliceerde artikelen van de afgelopen 5-10 jaar, die methoden bespreken om gepersonaliseerd nieuws mogelijk te maken, zie je dat er vele mogelijkheden voorhanden zijn. Voorbeelden zijn methoden gebaseerd op ontologieën [COT06] en kansberekening [ACT01] Ik zal onderzoeken wat de overeenkomsten en verschillen tussen deze methoden zijn. Daarnaast wil ik kijken welke informatie van gebruikers deze methoden gebruiken om hun voorkeuren te ontdekken. Hoewel er veel artikelen zijn die vooraf kort bespreken wat er al voorhanden is, voordat zij hun eigen creaties ten toon spreiden, ontbreekt een echt overzicht.

Om een indruk te krijgen van de werking van deze methoden in de praktijk, zal ik ook een case studie doen. Het nut hiervan is om behalve theoretisch allemaal methoden te behandelen, ook te kijken naar hoe 'grote' persoonlijke nieuwssites het aanpakken. Dit kan je veel vertellen over methoden die werken en methoden die niet werken.

2 Gepersonaliseerd Nieuws in de Literatuur



Figuur 1: De drie perspectieven en hun relaties

Om gepersonaliseerde nieuwssites mogelijk te maken, zijn er methoden nodig om profielen van gebruikers op te slaan, te bewerken en met deze profielen nieuwsberichten te filteren naar de voorkeur van de gebruiker. [MLM03] bekeken deze methoden vanuit twee perspectieven: ‘profile generation and maintenance’ en ‘profile exploitation’.

Het eerste perspectief valt uiteen in twee gedeeltes: ‘profile generation’ en ‘profile maintenance’. Samen met profile exploitation geeft dit drie perspectieven, waarmee hieronder personalisatiemethoden worden bekeken.

Deze drie perspectieven komen terug in figuur 1, een model waarin de samenhang tussen de verschillende elementen van een persoonlijk nieuwssysteem worden weergegeven.

Pijl 1 geeft de profile generation weer. Het systeem genereert eerst een gebruikers profiel. Hierna onderhoud het systeem het profiel door middel van interacties die de gebruiker met het systeem heeft. Dit is pijl 2, profile maintenance. Bij pijl 3 worden de nieuwsberichten geselecteerd en gepresenteerd op de interface van het systeem: profile exploitation. In de paragrafen hieronder wordt ieder van deze drie perspectieven verder uitgelegd.

Voor mijn literatuurstudie heb ik verschillende systemen bekeken. Hier volgt een kort overzicht van deze systemen en welke technieken en begrippen zij gebruiken. Deze technieken en begrippen worden in de onderstaande paragrafen toegelicht.

Tabel 1: Overzicht systemen uit de literatuur en hun gebruikte technieken

Systeem	KB	O	DF	IF	HDF	CF	CBF	K
Google News [DDG07]	x			x			x	x
Spotback	x		x			x	x	x
SEAN [ACT01]	x			x			x	x
PNS [COT06]		x						x
FAB [BS97]	x				x	x	x	x
ANTAGONOMY [SK97]			x				x	x
KRAKATOA [BKA98]			x	x				x
PNRM [LLK03]	x			x			x	x
Semantic Expansion [LYCK08]	x	x	x				x	x

KB = Kansberekening HDF = Half Directe Feedback
O = Ontologie CF = Collaborative Filtering
DF = Directe Feedback CBF = Content Based Filtering
IF = Indirecte Feedback K = Keywords

2.1 Profile Generation

Bij het opslaan en bijhouden van gebruikersprofielen zijn twee elementen van belang: de gegevens die de gebruiker expliciet opgeeft en gegevens die impliciet aan de gebruiker gekoppeld kunnen worden. Welke gegevens dit zijn, verschilt per gebruikte implementatie. Expliciete gevraagde gegevens zijn de gegevens die de gebruiker invult als hij zijn profiel aanmaakt, zoals naam, leeftijd en adres. Deze gegevens worden veelal gebruikt om een profiel te initialiseren. Maar op basis van alleen persoonsgegevens is het lastig om nieuws te filteren. Daarom worden ook gegevens als opleiding, werk en interesses bij de meeste systemen gevraagd [ACT00].

Het profiel wordt gebruikt als initialisatie van het systeem, welke later door het gedrag van de gebruiker aangepast en uitgebreid wordt. Bij alle methoden uit tabel 1 die ik heb bekeken, werd gebruik gemaakt van een expliciet ingevuld profiel. Daarnaast maakten de meeste systemen gebruik van gegevens die nodig waren om hun gekozen methode te laten werken. Een voorbeeld hiervan is SEAN. Dit persoonlijk nieuwssysteem, gecreëerd door Ardissono et al. [ACT00], koppelt aan een gebruiker de volgende vier dimensies:

Life Style Informatie over de prioriteiten en voorkeuren in levensstijl van de gebruiker

Interests Hierin staat in welke nieuwsonderwerpen de gebruiker is geïnteresseerd

Expertise Deze dimensie bevat informatie over de nieuwsonderwerpen waar de gebruiker 'expert' in is

Cognitive characteristics Informatie over de cognitieve grenzen die de gebruiker heeft, zodat het systeem kan bepalen hoeveel informatie de gebruiker van een bepaald detail-niveau

kan verwerken. Hieronder valt receptiviteit: de hoeveelheid informatie die een persoon kan lezen en verwerken.

Het invullen van deze dimensies gebeurt op basis van stereotypes. Aan iedere in te vullen waarde bij de standaard profielinformatie (naam, leeftijd etc.) wordt een kans gekoppeld. Bijvoorbeeld: bij het invullen van de interest dimensie, heeft iemand die werkzaam is in de financiële sector een kans van 1 dat hij is geïnteresseerd in economisch gerelateerd nieuws.

Op deze wijze kan m.b.v. weinig expliciet gevraagde informatie toch een uitgebreid gebruiker-profiel gecreëerd worden. Voor een persoonlijke nieuwssite is het erg belangrijk dat de gebruiker zo min mogelijk belast wordt met het uiten van zijn voorkeuren. De gebruiker wil immers zo snel mogelijk nieuws te zien krijgen, zonder eerst uren lang voorkeuren op te geven. Daarnaast is het voor gebruikers vaak moeilijk om in woorden aan te geven wat zijn voorkeuren nu precies zijn. Het probleem met stereotypes is wel dat het initiële profiel zo gedetailleerd is als dat er data van andere gebruikers voorhanden is [ACT00].

Behalve door gebruik te maken van kansen om een gebruikersprofiel aan te maken, kan men ook ontologieën gebruiken [MGB⁺04]. Als een gebruiker bepaalde interesses aangeeft, kan het systeem aan deze interesses gerelateerde onderwerpen activeren [COT06]. Hoe het profiel gegenereerd wordt, komt niet naar voren in het artikel van Conlan et al. [COT06]. Waarschijnlijk genereert het systeem een blanco ontologie, die naar mate de gebruiker feedback geeft (zie paragraaf hieronder) zichzelf langzaam uitbreidt. Alle technieken, zonder uitzondering, gebruiken keywords om de voorkeuren van hun gebruikers te weten te komen. De methoden die enkel en alleen keywords gebruiken, zien een gebruiker als een lijst van interesses en daaraan gekoppeld waardes (hoe interessant hij/zij iets vindt).

2.2 Profile Maintenance

De feedback die een gebruiker geeft op gevonden artikelen, om zo zijn profiel aan te passen en beter bruikbaar te maken voor het filteren van nieuws, wordt door Sakagami & Kamba [SK97] “*Somewhat indirect extraction of user preferences*” genoemd. Het is geen directe extractie, aangezien de gebruiker niet expliciet om zijn interesses wordt gevraagd. Maar het is ook weer geen indirecte extractie, want de gebruiker moet nog steeds wel zijn mening geven. Met deze feedback kan een systeem de waardes of keywords die in het profiel opgeslagen staan, aanpassen. Zo zal het SEAN-systeem de kansen aanpassen. Geeft een gebruiker dus aan dat hij een artikel over de politiek niet interessant vond, dan zal de kans gekoppeld aan politieke interesse in dit systeem verlaagd worden. De halfdirecte, halfindirecte feedback is voor een systeem erg belangrijk. Het is de meest effectieve manier om te weten te komen welk nieuws een gebruiker preferereert. Helaas heeft het grote nadelen: “*However, eliciting an explicit classification for all news articles a user reads is both intrusive and time-consuming. Classification tasks tend to cause undue cognitive loads and users tend to refrain from doing them. Furthermore, an explicit classification scheme would require awareness from the users that their interests are changing and there are no guarantees that the classification criteria will remain the same for all interactions with the system*” [CCGJ04]. Om deze redenen maken de door mij bekeken systemen er vaak optioneel gebruik van: het is voor de gebruiker mogelijk om nieuwsberichten te beoordelen, maar het is niet verplicht.

De laatste manier om een profiel van de gebruiker bij te houden, is de indirecte manier. Hierbij kijkt het systeem naar het gedrag van de gebruiker en extraheert hieruit de voorkeuren van de

2 Gepersonaliseerd Nieuws in de Literatuur

gebruiker. Bij alle systemen die werden genoemd in de literatuur, speelt indirecte feedback een grote rol. Ardissono et al. [ACT00] halen hiervoor de volgende redenen aan: *“First of all, the predictions made by the stereotypes cannot be very precise due to the limited set of data provided by the user in the registration form. Second, we are interested in providing a personalization for each individual user and thus we want to achieve predictions that are more accurate (individualized) than those provided by the stereotypes. Finally, and most important, the interests/priorities/goals of a user may change during time and we are interested in tracking these changes and modifying the user model accordingly.”* Een ander argument, wat ook al geldt voor de halfdirecte feedback, is dat een gebruiker vaak moeilijk zijn eigen voorkeur in woorden kan uitdrukken. Vaak is aan het gedrag van iemand beter te zien hoe hij is dan op grond van wat hij zelf zegt: *“Another problem is that people cannot necessarily specify what they are interested in because their interests are sometimes unconscious”* [SK97].

Bij de indirecte feedback zijn er verscheidene variabelen die van belang zijn. Neem SEAN, het systeem van Ardissono et al. [ACT00]. Zij gebruiken de volgende variabelen/events om het profiel aan te passen:

- De sectie en het nieuws dat de gebruiker verkent en de tijd die de gebruiker hier per sectie aan besteedt.
- De artikelen die het systeem heeft geselecteerd en de gebruiker wegklikte.
- De artikelen die het systeem niet selecteerde en de gebruiker wel las
- Het detailniveau van een nieuwsbericht, geselecteerd door het systeem, die de gebruiker wegklikte.
- De vraag van de gebruiker naar een hoger detailniveau dan het systeem selecteerde

Het systeem kijkt dus bijvoorbeeld welke onderwerpen door het systeem interessant werden gevonden, maar de gebruiker juist niet interessant vond. Door deze feedback worden dan de waardes (kansen) aangepast in de vier dimensies die het systeem gebruikt om nieuws te filteren. Hoe deze dimensies gecombineerd worden, wordt uitgelegd in de paragraaf hieronder.

Een andere benadering komt van het systeem van Carreira et al. [CCGJ04] en wordt ook gebruikt in het systeem van Lai et al. [LLK03]. Zij gebruiken vooral het element tijd om te meten hoe interessant de gebruiker een artikel vindt. De volgende variabelen worden gebruikt als indirecte feedback:

- Totale leestijd, in secondes (RT).
- Totaal aantal regels van artikel (NL).
- Aantal regels gelezen door gebruiker (NLR).
- K, de gemiddelde regelleestijd van de gebruiker.

Aan de hand van K wordt bepaald hoeveel regels de gebruiker van een artikel heeft gelezen. K wordt bepaald door de gebruiker een neutraal stukje tekst te laten lezen. Immers, iedere gebruiker heeft een andere leessnelheid. Dit heeft één groot nadeel, welke niet wordt aangehaald in het artikel van Lai et al. [LLK03]. Sakagami & Kamba [SK97] halen dit probleem wel aan:

“That is, not do other things such as leaving the terminal a while to get a cup of coffee or reading newly arrived e-mail messages. These conditions show the limitations of their method. In actual situations, we often receive e-mail and telephone calls and are subjected to other interruptions.” Dit betekent dus dat een gebruiker niet altijd aan het lezen is als een webpagina open staat. Hierdoor kunnen de metingen van Carreira et al. [CCGJ04] onbetrouwbaar worden.

Daarnaast kan men een click op een artikel zien als een positieve feedback [DDG07] en het niet klikken op een artikel zien als negatieve feedback [BKB⁺06].

Wat we dus zien in de literatuur is een onderverdeling in drie manieren om een profiel van een gebruiker aan te maken en te onderhouden: direct, halfdirect en indirect. De directe manier is vaak de enige optie om het initiële profiel aan te maken. Op basis van deze informatie bouwen de systemen een profiel, dat zij dan weer gebruiken om nieuws te filteren. Als er nieuws gepresenteerd wordt aan de gebruiker, gebruikt men halfdirecte en indirecte feedback om de voorkeuren van de gebruiker up-to-date te houden. Waar directe en halfdirecte feedback het nadeel hebben dat zij tijd kosten aan de gebruiker, zijn het vaak wel goede indicaties van de voorkeuren van gebruikers. Indirecte feedback kost de gebruiker daarentegen geen extra tijd, maar heeft als nadeel dat de metingen niet altijd even betrouwbaar zijn. Toch blijkt uit de artikelen dat de ervaringen met indirecte feedback positief zijn.

2.3 Profile exploitation

Nu we weten hoe profielen eruit zien en hoe ze bijgehouden worden, kunnen we kijken naar de volgende stap: het filteren van nieuws op basis van de profielen. Er zijn eigenlijk twee technieken te vinden op basis waarvan nieuws gefilterd wordt: content-based filtering en collaborative filtering [DDG07]. In het artikel van Liang et al. [LYCK08] worden content-based en collaborative filtering als volgt omschreven: *“Content-based filtering uses keywords or other product-related attributes to take recommendations. Collaborative filtering uses preferences of similar users in the same reference group as a basis for recommendation.”* Deze definitie gaat uit van de twee begrippen in algemene zin. De definitie gaat over hoe deze technieken worden toegepast in aanbevelingssystemen, het overkoepelende begrip waar gepersonaliseerd nieuws ondervalt. Een betere definitie, toegespitst op gepersonaliseerd nieuws komt van Das et al. [DDG07]. Zij zien content-based filtering als het zoeken naar overeenkomsten tussen bepaalde variabelen en aan de hand hiervan nieuwsberichten filteren. De variabelen zijn onderdelen van het profiel van de gebruiker en zijn ook terug te vinden in nieuwsberichten. Welke inhoud deze variabelen krijgen, verschilt per gebruikte methode. Maar in de meeste gevallen zullen dit keywords zijn. Collaborative filtering is volgens Das et al. [DDG07] het gebruiken van beoordelingen van overeenkomstige gebruikers voor nieuwsberichten en op basis van deze beoordeling het nieuws te filteren. De overeenkomst van de gebruikers wordt aan de hand van items in het gebruikersprofiel berekend.

Nadelen van content-based filteren zijn o.a. dat nieuwsberichten zich niet altijd makkelijk laten definiëren in bijvoorbeeld keywords [DDG07] en is het ook niet altijd mogelijk de items die zich in het gebruikersprofiel te vinden zijn te koppelen aan nieuwsberichten [LYCK08]. Dit zouden redenen kunnen zijn om over te stappen op een collaborative filtering systeem, maar ook zo'n systeem heeft nadelen. Als er nieuwe nieuwsberichten worden gepresenteerd, zijn deze lang niet allemaal beoordeeld door de gebruikers. Dit betekent dat het systeem moeilijk de nieuwe berichten kan aanbevelen aan hun gebruikers en dus dat gebruikers misschien erg interessant

2 Gepersonaliseerd Nieuws in de Literatuur

nieuws mislopen. Een ander nadeel is dat er een versterkend effect plaats vindt. Mensen zijn geneigd om berichten die het meest gelezen zijn, zelf ook te lezen. Nieuwsberichten die slecht gelezen of beoordeeld zijn, zullen hierdoor nog minder beoordeeld en gelezen worden, terwijl nieuwsberichten die erg goed gelezen of beoordeeld zijn juist nog beter beoordeeld en gelezen worden.

Als je naar exploitatiemethoden kijkt, zijn twee elementen altijd aanwezig: het ene element is de manier waarop zo'n methode de nieuwsberichten opslaat en representeert. Het andere element is hoe de methoden het gebruikersprofiel koppelt aan nieuwsberichten.

Zoals beschreven in de vorige paragraaf, zijn er verschillende principes die men kan gebruiken voor het achterhalen van voorkeuren van gebruikers. In alle artikelen die ik heb bekeken, gebruikte men hiervoor keywords en een daaraan gekoppelde waarde [LYCK08]. Daarnaast werd vaak een techniek als een ontologie of een probabilistische techniek zoals Bayes gebruikt. Het is ook van belang of een systeem zelf nieuwsberichten opslaat of dat het alleen linkt naar het nieuwsbericht op een andere site [DDG07].

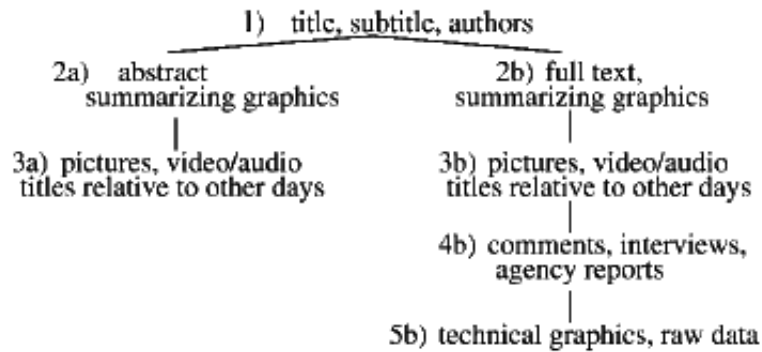
2.3.1 SEAN

SEAN, het systeem beschreven door Ardissono et al.[ACT00], maakt gebruik van het principe van kansberekening. Omdat dit systeem zo uitgebreid beschreven is, kan ik met behulp van dit systeem goed laten zien hoe overeenkomstige systemen mogelijk gebruik maken van kansberekening om nieuwsberichten te filteren. Bij SEAN wordt gebruik gemaakt van onderscheid in het detailniveau van nieuwsberichten. Nieuwsberichten die de gebruiker mogelijk minder interessant vindt worden bijvoorbeeld alleen met samenvatting en titel getoond. In hun artikel verdelen Ardissono et al.[ACT00] hun systeem op in drie onderdelen: een nieuwsdatabase, een advertentiedatabase en een gebruikersdatabase. Ik zal in deze paragraaf alleen de nieuwsdatabase behandelen. De advertentiedatabase valt buiten de scope van mijn onderzoek en de gebruikersdatabase heb ik besproken bij de paragraaf over profile generation.

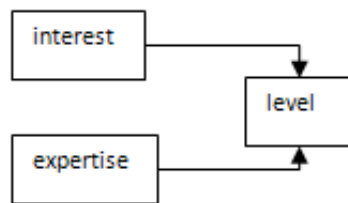
De nieuwsdatabase van SEAN moet twee mogelijkheden ondersteunen [ACT00]:

1. *“classifying news according to their topics and*
2. *generating different presentations (varying the detail level) of each news item.”*

Dit zijn echt specifieke eisen voor SEAN. Bij andere methodes zullen deze eisen anders zijn. Bij SEAN wordt gebruikt gemaakt van een verdeling van nieuwsberichten in topics en kan het detailniveau van nieuwsberichten worden aangepast. Het classificeren van nieuwsbericht op onderwerp, zo beschrijven zij, kan op twee manieren. *‘A first option is to classify news according to an a-priori hierarchy of topics. This is the choice made in most of the horizontal portals and in (traditional and electronic) newspapers, where the hierarchy has a correspondence in the structure of the editorial board. [...]. A second option is to generate the classification dynamically, using indices extracted from each news item. This is the technique adopted in most of the approaches to information retrieval’* [ACT00]. Zelf kiezen zij voor de eerste manier. Het is aan de beheerder van de nieuwsserver om de hiërarchie te creëren. De nieuwsberichten worden dan automatisch verdeeld over de secties. Zo heb je bijvoorbeeld de sectie politiek, met als subsecties nationale en internationale politiek. Daarnaast krijgt ieder bericht een detailniveau mee, dat beschreven wordt in figuur 2.



Figuur 2: Opbouw van artikelniveaus binnen SEAN



Figuur 3: Variabelen en hun relaties die het detailniveau bepalen

De personalisatie bij SEAN gebeurt aan de hand van twee agents:

1. *"An agent that personalizes the content of the presentation: given the pieces of information in the user model (i.e., information about the user interest, expertise, receptivity and life style), it decides which sections and news have to be presented, the appropriate detail level for each section and news item and the advertisements that have to be included in each page."*
2. *"An agent that generates the hypertextual pages. This agent could also be responsible for personalizing the form of the presentation"* [ACT00];

Voor mijn onderzoek is vooral de eerste agent interessant, dit is de agent die daadwerkelijk het nieuws filtert. De eerste stap voor deze agent is kijken wat de kans is dat een gebruiker van een bepaalde sectie met een bepaald detailniveau (beschreven in figuur 2) het nieuws wil lezen. Hiervoor wordt een matrix gebruikt die voor iedere sectie (S) opslaat wat de interesse van de gebruiker in die sectie is en welke expertise hij erin heeft. Er wordt dan uitgerekend wat de kans is dat, gegeven een interesseniveau en expertiseniveau, de gebruiker het nieuws met bepaald detail wil zien. De berekening is gebaseerd op de regel van Bayes [ACT00], met als variabelen interest, expertise en level (zie figuur 3). Waarbij "interest" en "expertise" de waarde van "level" bepalen.

Naast de regel van Bayes, is hier ook de combinatiefunctie voor kansen nodig [LvdG05]. Voorbeeld:

2 Gepersonaliseerd Nieuws in de Literatuur

$$P(\text{level} = 4 \text{ for } S | \text{interest in } S = \text{high, expertise in } S = \text{high}) = 0.2$$

$$P(\text{level} = 4 \text{ for } S | \text{interest in } S = \text{high, expertise in } S = \text{medium}) = 0.4$$

$$P(\text{level} = 4 \text{ for } S | \text{interest in } S = \text{medium, expertise in } S = \text{high}) = 0.4$$

$$P(\text{level} = 4 \text{ for } S | \text{interest in } S = \text{medium, expertise in } S = \text{medium}) = 0.7$$

Omdat interesse en expertise ook een kans hebben (een gebruiker heeft bijvoorbeeld een kans van 0.5 om een hoge interesse te hebben in een sectie), moet voor iedere mogelijk combinatie van interesse- en expertiseniveau de kans worden berekend.

Voorbeeld:

$$P(\text{interest in } s_i = \text{high}) = 0.2$$

$$P(\text{interest in } s_i = \text{medium}) = 0.7$$

$$P(\text{expertise in } s_i = \text{high}) = 0.3$$

$$P(\text{expertise in } s_i = \text{medium}) = 0.6$$

...

$$P(\text{level} = 4 \text{ for section } s_i) =$$

$$(P(\text{level} = 4 \text{ for section } S | \text{interest in } S = \text{high, expertise in } S = \text{high}) \times$$

$$P(\text{interest in } s_i = \text{high}) \times$$

$$P(\text{expertise in } s_i = \text{high}))$$

+

$$(P(\text{level} = 4 \text{ for section } S | \text{interest in } S = \text{medium, expertise in } S = \text{high}) \times$$

$$P(\text{interest in } s_i = \text{medium}) \times$$

$$P(\text{expertise in } s_i = \text{high}))$$

+

$$P(\text{level} = 4 \text{ for section } S | \text{interest in } S = \text{high, expertise in } S = \text{medium}) \times$$

$$P(\text{interest in } s_i = \text{high}) \times$$

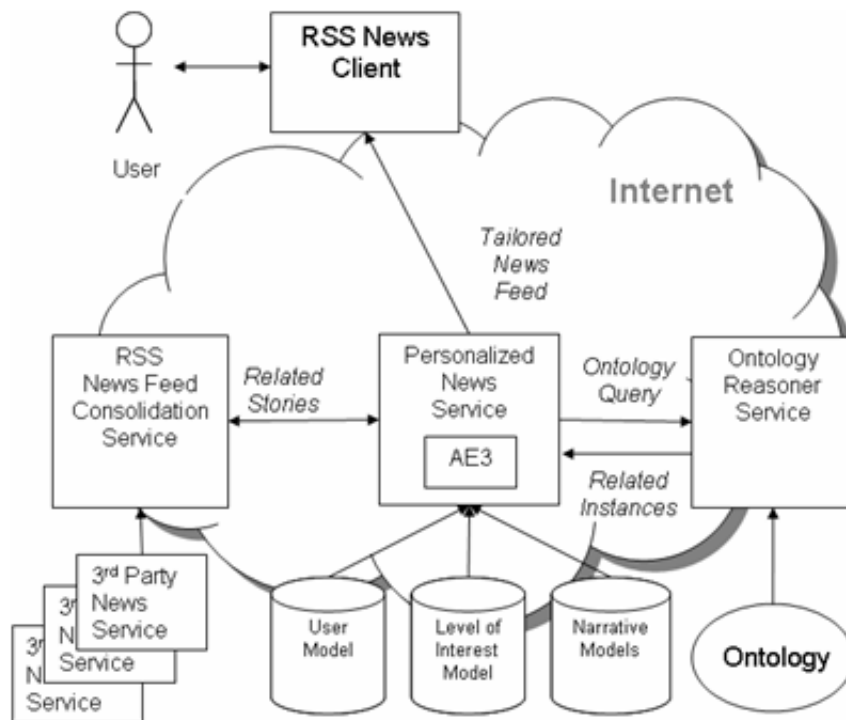
$$P(\text{expertise in } s_i = \text{medium}))$$

$$\begin{aligned}
& + \\
& (P(\text{level} = 4 \text{ for section } S | \text{interest in } S = \text{medium, expertise in } S = \text{medium}) \times \\
& \quad P(\text{interest in } s_i = \text{medium}) \times \\
& \quad P(\text{expertise in } s_i = \text{medium})) \\
& + \\
& \dots \\
& = 0.2 \times 0.2 \times 0.3 + 0.4 \times 0.7 \times 0.3 + \\
& + 0.4 \times 0.2 \times 0.6 + 0.7 \times 0.7 \times 0.6 + \dots
\end{aligned}$$

Je hebt nu de kans berekend dat iemand een bepaalde sectie s_i met detailniveau 4 wil zien. Deze berekening moet dus per detail niveau, per sectie worden gedaan om te weten hoe de gebruiker zijn nieuws wil zien. De volgende stap van de agent is om te bekijken welke secties gepresenteerd (de set P) gaan worden. Dit gaat met de volgende stappen:

1. alle secties waar detailniveau 0 (alleen titel en auteur, zie figuur 2) de hoogste score heeft, worden niet aan de set P toegevoegd.
2. alle secties waar de cumulatieve score van detailniveau 0 en 1 een kans hebben die boven 0.7 ligt, worden niet toegevoegd aan set P.
3. alle secties waar detailniveau 5 het hoogste scoort, worden aan set P toegevoegd
4. alle secties waar de cumulatieve score van detailniveau 4 en 5 boven 0.7 ligt, worden aan set P toegevoegd.
5. de overige secties worden geplaatst aan de hand van de verdeling van hun scores.

Nu moet bepaald worden hoeveel secties worden weergegeven en welk detailniveau iedere sectie krijgt. Eerst wordt aan de hand van de variabele receptiviteit (onderdeel van de dimensie Cognitive characteristics) bepaald hoeveel secties er worden weergegeven (het artikel vertelt helaas niet hoe dit gebeurt). Dan wordt gekeken per sectie S in de set P welk detailniveau van de sectie de hoogste score heeft. Er wordt dan gecontroleerd of dit detailniveau samengaat met het receptiviteitsniveau van de gebruiker (wederom wordt er niet verteld hoe dit gebeurt). Als dit samengaat, dan wordt S met het corresponderende detailniveau getoond. Is dit niet het geval, dan zijn er twee gevallen mogelijk. Het detailniveau kan te laag zijn in vergelijking met de receptiviteit die een gebruiker heeft. Er hoeft dan niks gedaan te worden. Er is immers berekend dat de gebruiker niet erg geïnteresseerd is in deze sectie. Mocht de gebruiker iets toch wel interessant vinden, dan kan hij het detailniveau altijd aanpassen. Is het detailniveau te hoog in vergelijking met de receptiviteit, dan zal het systeem een lager detailniveau moeten zoeken. Wederom wordt niet duidelijk hoe de agent (of het systeem) dit doet.



Figuur 4: PNS

2.3.2 Ontologieën

Met ontologieën werk het anders. Er zijn twee soorten ontologieën die gebruikt worden: ‘Strict’ en ‘loose’ ontologieën [COT06]. In hun artikel zetten Conlan et al. [COT06] wel hun vraagtekens bij de loose ontologieën. Waar de strikte nog betekenis geeft aan objecten en hun relaties, bevat de losse ontologie alleen maar objecten en relaties. Je zou het dan ook meer een gelinkte taxonomie kunnen noemen [COT06]. Het systeem van Conlan et al. [COT06], PNS, maakt daarnaast onderscheid tussen objecten en instanties van objecten. Als voorbeeld: een persoon kan fan zijn van Michael Schumacher. Hieraan gerelateerd vind je bijvoorbeeld Rubens Barrichello (zijn toenmalige teamgenoot). Het maakt een groot verschil of je gebruik maakt van de entiteit Barrichello of het object teamgenoot. Teamgenoten kunnen immers veranderen, terwijl het nieuws omtrent Barrichello altijd erg specifiek zal zijn. Als expliciete informatie moet een gebruiker wel aangegeven hoe erg zijn voorkeuren uitgaan naar de door het systeem gecreëerde, gebruiker specifieke ontologie. De mogelijke waarden van voorkeur zijn: none, low, medium and high. Uit het artikel valt op te maken dat hoe hoger de waarde van voorkeur ligt, hoe meer het systeem uit de ontologie gebruik maakt van entiteiten i.p.v. objecten.

In het artikel van Conlan et al. [COT06] wordt een systeemmodel gegeven dat goed laat zien hoe hun systeem werkt.

Het systeem van Conlan et al. [COT06] is bedoeld voor RSS-feeds, op bijvoorbeeld PDA’s. Het systeem bestaat uit 3 hoofdcomponenten:

1. RSS News Feed Consolidation Service (CS)

2. Personalized News Service (PNS)
3. Ontology Reasoner Service (ORS)

De CS registreert nieuws berichten van derden, bijvoorbeeld nieuwsberichten van commerciële nieuwskanalen. Daarnaast voert de CS keyword-queries uit op deze nieuwsberichten. De ORS is de ontologie, die het mogelijk maakt om met queries naar relaties te zoeken.

Het hart van het systeem is de PNS. Deze stelt een query op aan de hand van de interesses van de gebruiker. Deze query wordt dan verstuurd naar de ORS. Aan de hand van query wordt er in de ontologie gezocht naar onderwerpen die mogelijk ook van interesse zijn voor de gebruiker. Deze onderwerpen worden teruggestuurd naar de PNS. Deze stuurt de onderwerpen door naar de CS, die een query uitvoert van keywords op de nieuwsberichten. Deze nieuwsberichten worden dan teruggestuurd naar de PNS en vanuit daar naar de RSS News Client.

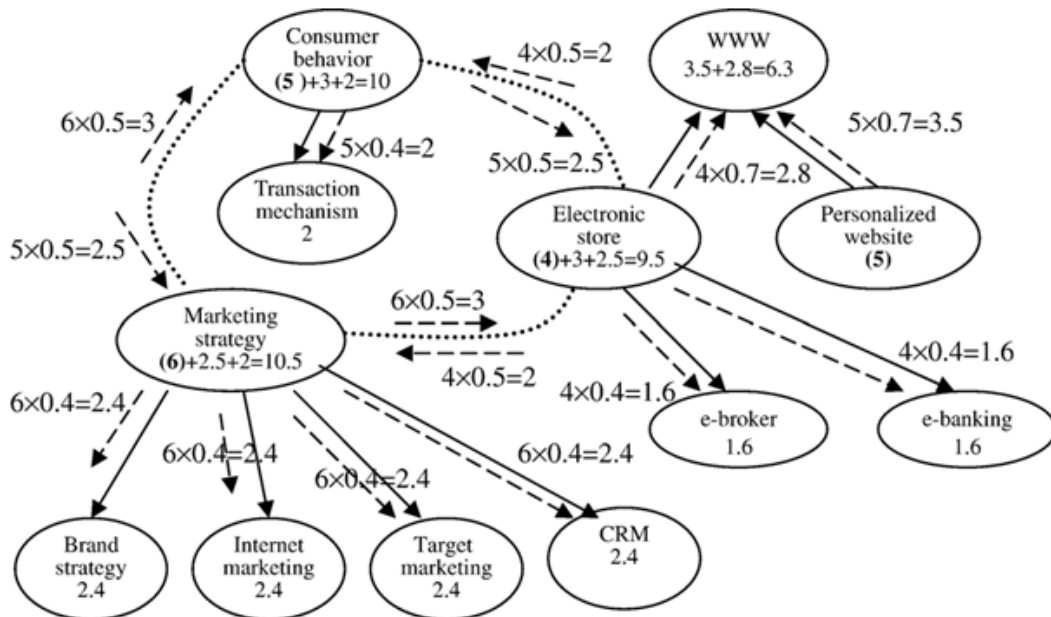
Helaas is in het artikel van Conlan et al. [COT06] niet terug te vinden hoe zij al deze stappen uitvoeren. Zij noemen wel technieken en talen, zoals WSDL en RDQL, maar de echte implementatie wordt niet genoemd. Hierdoor klinkt dit allemaal wel prachtig, maar is het de vraag of het ook allemaal werkt zoals zij zelf zeggen. Hun eigen evaluatie, het laten evalueren van de service door 10 gebruikers, bekeek of er verschil zat tussen ‘loose’ en ‘strict’ technieken voor ontologieën. Daarnaast gebruikten zij hiervoor een zeer beperkte ontologie (alleen maar Formule 1 nieuws). Er is dus niet getest of de server ook wel echt persoonlijk nieuws oplevert. Hoewel deze techniek dus leuk klinkt in theorie, wordt de werking ervan naar mijn mening niet overtuigend aangetoond.

Een artikel dat wel de werking beschrijft van een soort ontologie, is het artikel van Lai et al. [LLK03]. Zij maken gebruik van semantische bomen en netwerk. Een semantische boom lijkt erg veel op een ontologie en kan in een semantisch netwerk aan andere semantische bomen worden gekoppeld. In het semantisch netwerk staan allerlei concepten met relaties daartussen (zie figuur 5).

In het netwerk vind je twee soorten relaties: *“The solid arrow shows an is-a relationship that implies certain property inheritance. The dashed line shows a non-is-a relationship, i.e., all other possible associations between two semantic trees with no property inheritance”* [LYCK08]. Je ziet hier dus dat Consumer Behavior, Marketing Strategy en Electronic Store de drie semantische bomen zijn, gekoppeld door een non-is-a relatie. De onderbroken pijlen geven alleen de richting aan van de berekening van de waardes, welke zometeen aan bod komen.

In een semantische netwerk zijn er dan weer drie verschillende relaties te vinden: generalisatie, specialisatie en relevantie expansie. Generalisatie gaat over het activeren van een concept boven het huidige concept (bijvoorbeeld de relatie Personalized Website \rightarrow WWW), specialisatie over het activeren van een concept onder het huidige concept (bijvoorbeeld Electronic Store \rightarrow e-broker). Relevantie expansie is het activeren via een non-is-a relatie (dus tussen twee semantische bomen, bijvoorbeeld Consumer Behavior en Marketing Strategy).

Voor de duidelijkheid zal ik een voorbeeld doen hoe op basis van deze techniek nieuws gefilterd kan worden. De initiële basis interesses van een profiel zijn hier een vector: [consumer behavior (5), electronic store (4), Personalized website (5), marketing strategy (6)]. De waardes van deze drie concepten worden dus geïnitieerd op 5, 4 en 6. Het systeem geeft ook waardes aan de drie verschillende relaties. Generalisatie krijgt een waarde van 0.7, specialisatie van 0.4 en relevantie van 0.5. Om te bepalen welke concepten gebruikt gaan worden en welke niet, wordt er een threshold-value ingesteld van 2.4 (hoe deze bepaald wordt, is niet duidelijk). Zoals in



Figuur 5: Semantisch

het semantisch netwerk is te zien, wordt op basis van deze waarden, de waarden uitgerekend voor de andere concepten. Als je de waarde van een concept een niveau lager dan bijvoorbeeld Internet Marketing wil uitrekenen, gebruikt het systeem dus weer de waardes die de drie soorten relaties hebben gekregen en de waarde die net is uitgerekend: 2.4. Met de threshold worden nu de oninteressante concepten weggelaten. Dit zijn bijvoorbeeld e-broker (waarde 1.6) en Transaction Mechanism (waarde 2.0).

Het filteren van artikelen gaat nu als volgt: er wordt van ieder nieuwsartikel een keyword-vector opgesteld. Door middel van een synoniemenlijst wordt deze vector gekoppeld aan de concepten in het semantisch netwerk. Er wordt dan een optelsom gedaan van alle verschillende waardes van de concepten die in de vector zitten. De nieuwsartikelen waarvan de vector de hoogste waarde heeft, wordt als eerste getoond, etc.: *“For example, if a document has three keywords [Electronic CRM, Data Mining, and marketing strategy], then we can find that the interest value of these keywords to the user are 2.4, 0 and 10.5, respectively. Its total interest value is 12.9. The system can then make recommendations according to their interest values of the analyzed documents”* [LYCK08]. In het artikel van Liang et al. [LYCK08] wordt, in tegenstelling tot bij Conlan et al. [COT06] wel aangetoond door hun case study dat hun mechanisme werkt: *“We can see that the articles recommended by the semantic-expansion approach better caught user interests, compared with the keyword approach (5.18 vs. 5.03 on a 7-point scale)”*.

2.4 Samenvattend

Voor het filteren van nieuws zijn er twee algemene technieken die gebruikt worden: in de meeste gevallen wordt er gebruik gemaakt van kansrekening, in sommige gevallen van ontologieën/semantische netwerken. Ik heb laten zien hoe op basis van deze technieken nieuws gefil-

terd kan worden. Op basis van wat ik in de literatuur zie, is kansberekening wel de meest vooraanstaande techniek om te gebruiken. Ontologieën/semantische netwerken worden weinig gebruikt, maar een echte reden is in de literatuur niet te vinden. Een verklaring hiervoor kan gezocht worden in het feit dat systemen die gebruik maken van ontologieën/semantische netwerken ingewikkelder zijn dan systemen die gebruik maken van kansberekening. Er moet veel tijd gestoken worden in het genereren van een ontologieën/semantische netwerken voor een gebruiker en nog meer tijd in het bijhouden van de relaties tussen de concepten van de ontologieën/semantische netwerken en hun waardes. Kansberekening biedt voor persoonlijke nieuwssites een snellere, makkelijkere en, na bestudering van de literatuur, beter werkende oplossing.

3 Case Studie

3.1 Uitleg

Voor de case studie bekijk ik twee persoonlijke nieuwssites, Google News [New09a] en Spotback News [New09b]. Ik heb per site twee mensen deze een maand lang laten gebruiken: RJ & JJ hebben Google News bekeken, CR & PB Spotback News. Zij gebruikten de sites als nieuwssite: ze zochten nieuws, lazen artikelen en beoordelen waar mogelijk de interessantheid van deze artikelen. Iedere sessie hielden zij bij hoeveel artikelen zij lazen en hoeveel er daarvan echt interessant waren voor hen.

Om de twee sites beter te leren begrijpen, zal ik hieronder eerst beide sites bespreken. Ik vertel kort de geschiedenis van de sites, hoe zij werken en welke technieken uit de vorige twee paragrafen zij gebruiken om persoonlijk nieuws aan te bieden. Daarna vertel ik gedetailleerder over hoe ik mijn case studie heb uitgevoerd en welke resultaten ik hieruit heb gekregen. Maar we beginnen met uitleg over Google News en Spotback News.

3.1.1 GoogleNews

Het internet Afbeeldingen Video Maps Nieuws Discussiegroepen Gmail meer ▾ k.hulsman@gmail.com | Webgeschiedenis | Mijn account | Afmelden

Google nieuws Nederland Nieuws zoeken Het web doorzoeken Geavanceerd zoeken naar nieuws Voorkeuren

Aangepast nieuws ▾ Voropaginanieuws Bijgewerkt: 9 minuten geleden

Voorpaginanieuws

09:25 Oprolbaar asfalt op lange baan
De Telegraaf - 49 minuten geleden
DEN HAAG - Rijkswaterstaat staakt een proef met oprolbaar asfalt. Het product levert vooralsnog niet de gewenste resultaten op. Sinds de proef circa drie jaar geleden begon, zijn drie testvakken aangelegd met Rollpave, het nieuwe asfalt. ...
Prof 'asfalt op de rol' mislukt | Blik op Nieuws
Prof met afrolbaar asfalt mislukt | NOS.nl
BN/De Stem - Nijmegen nieuws - Infraside - RTV Oost
[alle 23 nieuwsartikelen »](#) [E-mail dit artikel](#)

Centrum Kinderontvoeringen: Katja hoort in Ede
Gelderlander - 49 minuten geleden
EDE - De Verenigde Staten dienen Katja Leendertz uit te wijzen naar Nederland. Het ontvoerde meisje mag dan de Amerikaanse nationaliteit bezitten, ze hoort thuis in Nederland. Door zijn 7-jarige dochter zomaar mee naar Amerika te nemen pleegde vader ...
Moeder ontvoerde Katja mishandeld | De Telegraaf
Vragen aan ministers over Katja | Ede Stad
Volkskrant - AD.nl - Omroepgelderland - Midland FM
[alle 88 nieuwsartikelen »](#) [E-mail dit artikel](#)

Aanbevolen voor k.hulsman@gmail.com » [Meer informatie](#)

'Micky laat mij leven'
De Telegraaf - 3 uren geleden - AMSTERDAM - Adam Curry is door Micky Hoogendijk een ander mens geworden. „Ik zag grijs, maar nu zie ik kleuren.“ Dat ...
Primeurjagers.nl - AD.nl - NU.nl
[alle 27 nieuwsartikelen »](#)

Wrakstukken van Airbus Air France
BN/De Stem - 1 uur geleden - PARIJS (ANP) - Wrakstukken die in de Atlantische Oceaan zijn gevonden bij de zoektocht naar de verdwenen Franse Airbus ...
Brabants Dagblad - Knack - AD.nl
[alle 497 nieuwsartikelen »](#)

Dag Jacqui
DePers.nl - 2 uren geleden - Gesjoemel met een tweede huis, twee pornofilms voor haar man en een badstop van 88 pence deden haar de das om. De Britse ...
Volkskrant - De Telegraaf - Elsevier
[alle 24 nieuwsartikelen »](#)

Chinezen nemen Hummer over
Elsevier - 1 uur geleden - [alle 83 artikelen »](#)

Google gaat boeken verkopen
De Telegraaf - 18 uren geleden - [alle 14 artikelen »](#)

Twente-voorzitter Munsterman is zeker van komst Cissé
Sportweek - 41 minuten geleden - [alle 60 artikelen »](#)

Pinkpop viert verjaardag met een gedegen feest
NRC Handelsblad - 19 uren geleden - [alle 127 artikelen »](#)

Vierde geval Mexicaanse griep in Nederland
Nieuws.nl - 20 uren geleden - [alle 32 artikelen »](#)

Peking op scherp om herdenking
De Telegraaf - 1 uur geleden - [alle 86 artikelen »](#)

Pastor hekelt rol media in familiedrama
BN/De Stem - 2 uren geleden - [alle 65 artikelen »](#)

In het nieuws

Roda JC	Xbox 360
Super Mario	Wii Fit
Final Fantasy	Edward Sturing
Cambuur Leeuwarden	Geert Wilders
Churandy Martina	De Graafschap

Buitenland » [bewerken](#) **Nederland »** [bewerken](#)

Peking op scherp om herdenking [bewerken](#)
De Telegraaf - 1 uur geleden

Nederlandse schepen zoeken naar Airbus [bewerken](#)
Elsevier - 42 minuten geleden

Figuur 6: Screenshot Google News

Google News is een van de vele services die Google online aanbiedt. Zo kan je via Google gebruik maken van een emailaccount, weerberichten, tekstverwerking en online videos. Wat Google News onderscheidt van de andere services is dat het de eerste service van Google was die zich richtte op personalisering. Hoewel Google News tegenwoordig onderdeel lijkt van iGoogle, is dit niet zo. iGoogle is een soort internetstartpagina, waar gebruikers door middel van gadgets (kleine venstertjes) allemaal verschillende Google services op 1 pagina kunnen krijgen en is in mei 2005 online gekomen.

Google News daarentegen wordt al sinds 2001 ontwikkeld. Het heeft een hele lange tijd in een testfase gedraaid en is sinds 2006 officieel uitgebracht. Het bestond dus al veel eerder dan iGoogle. Volgens de statistieken heeft Google News per week ongeveer een miljoen unieke gebruikers [DDG07]. Net als bij iGoogle is het hoofdscherm van Google News naar voorkeur aan te passen. Je kunt aangeven welke nieuwssectie hoeveel berichten per nieuwssectie er worden weergegeven, hoe deze over de pagina verdeeld staan en e-mail notificaties laten versturen als

3 Case Studie

er berichten verschijnen met bepaalde keywords, die jij hebt aangegeven. Het doel van Google News is het bieden van meer persoonlijke opties en een breder perspectief [New09a]. Google probeert per nieuwsartikel meerdere links te bieden naar verschillende nieuwssites, zodat een gebruiker zelf kan kiezen van welke bron hij/zij het bericht leest.

De reden dat Google News interessant is voor mijn onderzoek, is de gadget ‘aanbevolen voor’ (zie figuur 6) op de hoofdsite van Google News. Hier staan nieuwsberichten die op basis van de voorkeuren van de gebruiker worden getoond. Waar de gehele site van Google News door de gebruiker aangepast kan worden, verzorgt Google bij de ‘aanbevolen voor’ sectie voor de aanpassing aan de gebruiker. Zij proberen door middel van het surfgedrag van de gebruiker te bepalen welke artikelen mogelijk interessant zijn.

Om te bepalen welke artikelen mogelijk interessant zijn, moet Google News eerst een profiel aanmaken voor een gebruiker. Google gebruikt hiervoor je Google-account: een account dat je toegang biedt tot alle online services van Google (bv. Gmail). Google News gebruikt alleen maar impliciete feedback om je voorkeuren te leren kennen. Op je Google account staat allemaal informatie opgeslagen, waaronder je zoek- en klikgeschiedenis bij bijvoorbeeld Google Search. Het profiel, waarop het filteren wordt gebaseerd, bevat keywords; termen waar je ooit op hebt gezocht of op hebt geklikt. Daarnaast kijkt Google News naar de nieuwsberichten waar je op klikt [New09a] Dit zien zij als een positieve feedback en de keywords van dat artikel krijgen een hogere waardering in je profiel [DDG07].

Natuurlijk is deze methode erg foutgevoelig. Gebruikers klikken vaak genoeg op artikelen die zij totaal niet interessant vinden. Toch gaat de rating voor de gerelateerde keywords dan wel omhoog. Andere methoden die in paragrafen 2.2 wordt genoemd, zoals hoe lang een gebruiker een artikel leest, worden door Google News niet gebruikt. Nadat je op een artikel hebt geklikt, wordt je doorgestuurd naar de corresponderende nieuwsbron en verlaat je de pagina van Google News.

Over het filteren van nieuws op basis van je profiel is niet veel te vinden. Toch valt er op basis van wat we weten wel op te maken hoe het werkt. Aangezien Google News gebruik maakt van keywords en gekoppelde ratings, zoekt de service in de gekoppelde nieuwsbronnen (zo’n 4500 internationale bronnen, waarvan 400 Nederlands) naar keywords die zich in het gebruikersprofiel bevinden. Artikelen die keywords bevatten met de hoogste rating in het profiel, zullen als eerste gepresenteerd worden. Google gebruikt dus content-based filtering om artikelen te vinden. Hoewel er geen expliciete bronnen zijn die vermelden welke algemene methode Google News gebruikt, lijkt het erg overeen te komen met de kansberekeningssystemen uit de literatuur (zie tabel 1). Hierdoor acht ik het aannemelijk dat zij kansberekening gebruiken als algemene methode om nieuws te personaliseren.

Naast het filteren op basis van de voorkeur van de gebruiker, probeert Google News ieder artikel ook meerdere links te geven naar verschillende nieuwsbronnen. Een bepaalde gebeurtenis wordt meestal in meerdere nieuwsbronnen uiteengezet en het is fijn als de gebruiker zelf kan kiezen van welke bron het bericht wordt gelezen. Dit strookt ook weer met het doel van Google News: *“We hebben ons ten doel gesteld onze lezers meer persoonlijke opties en een breder perspectief te bieden”* [New09a]. Google News is dus een persoonlijke nieuwssite, die wekelijks door miljoenen wordt gebruikt. Het probeert om op 1 pagina een variëteit aan nieuws te presenteren, aanpasbaar voor de gebruiker. Daarnaast probeert het, d.m.v. impliciete feedback, nieuws aan te bevelen aan de gebruikers, die past bij hun interesses.

3.1.2 Spotback News

The screenshot shows the Spotback News website interface. At the top, there is a navigation bar with the Spotback logo and a "get started" button. Below the navigation bar, there are various category links such as "Sports", "Computers and Internet", "Technology", "Business", "Science", "Entertainment", "General News", "Arts", "Health", "Hobbies", "Children", "Women", "Education", "Blogs", "Games", "Professional", "Music", and "Track specific keywords".

The main content area is divided into several sections:

- Computers and Internet:** A story titled "Why is Sequoia Looking into Associative Browsing Add-on SimilarWeb?" by TechCrunch, published yesterday. The story discusses Yahoo's purchase of FoxyTunes and its implications for Israeli startups.
- Technology:** A story titled "Yondi - cool travel pillow for kids" by The Red Ferret Journal, published 2 hours ago. The story describes the Yondi Travel Pillow, designed to prevent head flopping in children.
- Mobile Phones:** A story titled "Microsoft Bing: 7 Quick 'n' Dirty Tricks [Bing]" by Gizmodo, published yesterday. The story lists seven "dirty tricks" of the Bing search engine.
- Games:** A story titled "E3 2009: Transformers: Revenge of the Fallen Videos" by IGN Complete, published 2 days ago. The story reports on new footage of big robots from the movie.
- General News:** A story titled "Obama tells Germany of Middle East confidence" by The Guardian World News, published 22 hours ago. The story reports on President Obama's speech in Cairo and his meeting with Angela Merkel.

Each story includes a title, author, date, and a "Rate this story" widget. The "Rate this story" widget shows a progress bar and a "0" rating.

Figuur 7: Screenshot Spotback News

Daar waar Google News onderdeel is van het grote Google, is Spotback News een op zichzelf staande site. Een echte geschiedenis van Spotback is niet te vinden. Er is zelfs geen begindatum te vinden van wanneer de service online is gekomen. Op hun eigen blog zien we dat zij op 1 mei 2006 officieel online zijn gegaan [Blo09]. Tot 2007 was Spotback puur en alleen een site om nieuws te raten. Maar vanaf 2007 bieden zij een applicatie aan die het mogelijk maakt om, zoals ze zelf zeggen, alles te raten (de ondertitel van spotback.com is rate everything). Het is hierbij vooral de bedoeling dingen als websites, muziek en dus ook nieuws te raten.

Spotback's gespecialiseerde nieuwssite, news.spotback.com, geeft net als Google News een overzicht in allerlei categorieën: financieel nieuws, technisch nieuws, sportnieuws en dergelijke.

3 Case Studie

Het grote verschil met Google is dat Spotback alleen maar Engelstalige sites gebruikt om nieuws te vinden en dit vaak ook de wat kleinere nieuwsbronnen zijn (dus geen bronnen als New York Times e.d.). Er is, helaas, dus geen Nederlandstalig nieuws te vinden.

Spotback News biedt wel de mogelijkheid om de site naar hartelust aan te passen. Vensters kunnen anders geplaatst worden, categorieën kunnen worden toegevoegd en kleuren worden aangepast. En bovenal: het nieuws past zich aan als jij nieuwsberichten ‘rate’; dat wil zeggen: beoordeelt.

Bij ieder nieuwsbericht is een beoordelingsbalkje te vinden, dat normaal in het midden staat. Beweeg je het balkje naar links, dan geef je een artikel een negatieve beoordeling. Beweeg je het naar rechts, dan beoordeel je het nieuws positief. Op basis van deze feedback filtert Spotback de nieuwsberichten opnieuw en toont het de resultaten.

Om gebruik te maken van feedback, moet je eerst een profiel aanmaken. Deze bestaat eigenlijk alleen uit basisgegevens: een emailadres, gebruikersnaam, wachtwoord. Daarnaast kan je je eigen website nog invullen en een hip profielplaatje (ook wel avatar genoemd) uploaden. Spotback genereert dan een ‘default profile’, een gemiddeld profiel van alle gebruikers. Op basis hiervan wordt de eerste keer de site geladen. Je kunt dus zeggen dat de eerste keren er vooral gebruik gemaakt wordt van collaborative filtering: nieuws wordt gefilterd op basis van wat anderen beoordelen. Naar mate je de site vaker gebruikt en veel artikelen beoordeelt, gaat het zich steeds meer aan jouw voorkeuren aanpassen.

Spotback maakt duidelijk gebruik van directe feedback: je moet zelf de artikelen beoordelen om het algoritme te trainen. Omdat ook over Spotback weinig bronnen te vinden zijn, is het gokken met welke methode berichten gefilterd worden. Duidelijk is wel dat een positieve beoordeling van een artikel bepaalde keywords een plus geven. Waarschijnlijk worden keywords met de meeste plussen het belangrijkste gevonden en het eerst opgezocht.

3.1.3 Overeenkomsten

Zowel Google als Spotback maken dus gebruik van keyword-rating: bij allebei de sites krijgen de gerelateerde keywords bij positieve feedback een grotere positieve waarde. Helaas bestaat bij Google geen negatieve feedback, terwijl deze bij Spotback wel duidelijk aanwezig is. Natuurlijk komt dit ook door het feit dat bij impliciete feedback, waar Google gebruik van maakt, negatieve feedback een stuk moeilijker te meten is dan positieve feedback. Spotback, met de expliciete feedback, heeft het daarin een stuk makkelijker. Maar zoals eerder beschreven vraagt dit wel meer van de gebruiker. Je moet dus echt zelf moeite in de site steken, anders gebeurt er ook niks. Bij Google is alleen nieuwsberichten aanklikken genoeg.

Beide sites hebben dus, in vergelijking met elkaar, iets interessants te bieden. Google heeft impliciete feedback en daarnaast veel meer gebruikers dan Spotback, met zijn expliciete feedback en vaak kleinere bronnen. Voor de case studie is dit een leuke vergelijking. Hieronder zal ik het idee van mijn case studie en de resultaten uiteenzetten.

3.2 Case Studie Beschrijving

Met de literatuurstudie als uitgangspunt, heb ik ook een case studie gedaan. De bedoeling van de case studie is om te bekijken of persoonlijke nieuwssites ook echt werken. Filteren zij werkelijk het nieuws zo dat het aanbevolen nieuws voor jou ook echt interessant is voor je? Hoe lang duurt het voor dat er echt resultaat te vinden is?

Om deze vragen te beantwoorden heb ik vier mensen gevraagd of zij Spotback News en Google News een maandlang wilden gebruiken. Twee gebruikten Spotback en twee Google als nieuwssite. Voor allebei de sites hielden zij bij hoeveel artikelen zij lazen. Bij Google hielden zij daarnaast bij hoeveel van deze artikelen ‘interessant’ bevonden werden; interessant moet hier gezien worden als een artikel dat inspeelt op de persoonlijke interesse van de gebruiker.

Bij Spotback hielden de testpersonen bij zij positief hebben beoordeeld. Een positieve beoordeling wordt dan als interessant gezien. Hieruit valt ook weer te abstraheren hoeveel zij er niet interessant vonden: totaal aantal gelezen nieuwsberichten - aantal interessante nieuwsberichten = aantal oninteressante nieuwsberichten. Het idee achter het bijhouden van aantal gelezen artikelen en het aantal interessante artikelen is als volgt: na verloop van tijd zal een algoritme zich aan moeten passen aan de voorkeuren van de gebruiker. Dit betekent dat het aantal interessante artikelen zal moeten toenemen.

Natuurlijk is het sterk afhankelijk van het aangeboden nieuws hoe interessant iets is. Er is niet altijd interessant nieuws te bieden, hoe goed een algoritme ook werkt. Toch kun je zeggen dat, hoewel er dus mindere dagen tussen zitten, er een stijgende lijn in het aantal interessante artikelen moet zitten om te kunnen spreken van een succesvolle persoonlijke nieuwssites.

Het belangrijkste punt in mijn case studie is het meten van de kwaliteit van persoonlijke nieuwssites. Om deze kwaliteit te meten is er één variabele erg belangrijk: de gelezen nieuwsberichten. De andere variabelen die belangrijk zijn, zoals kwaliteit van bronnen, worden in de evaluatie-enquête getest.

In de evaluatie-enquête wil ik nog achter wat andere variabelen komen. Dit zijn de variabelen interactie en kwaliteit & variëteit van de nieuwsbronnen. Samen met de variabele personalisatie bepalen zij de kwaliteit van de persoonlijke nieuwssite. De overzichtelijkheid wordt getoetst in vraag 1, de gebruikersvriendelijkheid (snelheid, kan ik overal makkelijk op klikken?) in vraag 2. Deze twee variabelen bepalen hoe goed de interactie met de gebruiker is via de website. De bronnen worden beoordeeld in vraag 3 (aanbod) & 4 (betrouwbaarheid). Betrouwbare bronnen kun je zien als algemene aanvaarde nieuwsbronnen, zoals die van online kranten (Telegraaf, Volkskrant) en nieuwssites (Nu.nl, Nos.nl) Vraag 5 is een vraag om te toetsen of de gebruiker het idee had dat de site echt iets deed aan personalisering. Als dit niet zo is, dan betekent het dat de site slecht werk heeft afgeleverd. Immers, het doel van een persoonlijke nieuwssite is om de gebruiker het idee te geven dat het nieuws gepersonaliseerd wordt. Vraag 6 & 7 zijn bedoeld als toetsing van de attitude van de gebruiker met betrekking tot de nieuwssite. Hiermee probeer ik te achterhalen wat er concreet mis of juist heel goed is aan de site.

Ik hoop te ontdekken dat er een stijgende trend zit in het aantal interessante artikelen dat wordt aangeboden. Met deze uitkomst kan ik zeggen dat het algoritme in ieder geval in staat is om de voorkeuren van gebruikers te leren.

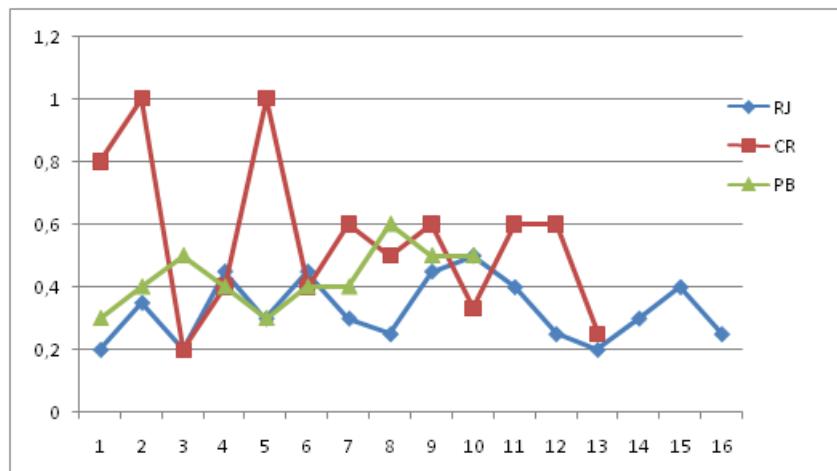
Als er juist een gelijke of dalende trend in zit, kun je zeggen dat het algoritme zijn werk niet goed doet. Je haalt niet het resultaat uit de site die eigenlijk beloofd wordt; persoonlijk nieuws.

3.3 Resultaten

Van de vier testpersonen, zijn er drie in staat geweest regelmatig de twee sites te bekijken en het aantal artikelen te noteren. Hun resultaten zijn weergegeven in figuur 8.

Te zien valt hier de interessantheids-index (aantal interessante artikelen / aantal gelezen artikelen). Waar ik verwachtte in deze grafiek een stijgende lijn te zien (naar mate het systeem

3 Case Studie



Figuur 8: Resultaten Case Studie

vaker gebruikt zou worden, zou het aantal interessante artikelen moeten toenemen) zien we dat hier geen sprake van is. Bij alle drie mijn testpersonen blijft de lijn met pieken en dalen voorwaarts gaan. Er is uit deze grafiek in ieder geval geen trend opwaarts te ontdekken.

Maakt dit allebei de websites slechte persoonlijke nieuwswebsites? Nee, dat hoeft niet. Er is een goede verklaring te vinden waarom de resultaten niet zijn wat ik er van verwachtte, maar de sites nog steeds goed zijn in persoonlijk nieuws aanbieden. In drie van de vier evaluaties valt te lezen dat de testpersonen het nieuws eenzijdig vinden worden, naarmate zij de site meer gebruiken: *“Als je bv. twee artikelen die over Apple gaan positief rate wordt je daarna meteen overspoeld met Apple nieuws. Dat wil je dan ook weer niet. Ik ben vooral geïnteresseerd in een gezonde nieuwsmix, beetje sport, politiek, buitenland nieuws etc.”* (PB). Het probleem met de twee websites is dus niet dat de personalisatie niet goed werkt, het probleem is dat hij te goed werkt. Omdat er in het begin weinig informatie is over de interesses van de gebruiker, wordt er veel hetzelfde nieuws weergegeven. Omdat ander nieuws minder aangeboden wordt, gaat de gebruiker juist weer op hetzelfde nieuws klikken, waardoor deze extra versterkt wordt (het versterkende effect uit de paragraaf 2.3).

Het wordt duidelijk uit de evaluaties dat gebruikers vooral geïnteresseerd zijn in een nieuwsmix, en niet zozeer in nieuws specifiek aan één onderwerp. Ik zal nu de evaluaties per nieuwssite bespreken.

3.3.1 Google News

Tabel 2: Resultaten Enquête Google News

Vraag	Antwoord RJ	Antwoord JJ
Overzichtelijkheid	Erg goed	Goed
Gebruiksvriendelijkheid	Goed	Goed
Nieuwsaanbod	Goed	Erg Goed
Betrouwbaarheid Bronnen	Betrouwbaar	Geen mening
Relatie nieuws aan eerder bekeken nieuws	Ja	Soms wel, soms niet
Ga je de site vaker gebruiken?	Ja	Ja
Waarom vaker gebruiken?	<ul style="list-style-type: none"> • Breed scala aan nieuws • Overzichtelijk • Leest net zo makkelijk als een krant 	<ul style="list-style-type: none"> • Overzichtelijk • Bevat email-account • Snel toegang tot nieuws van mijn studie (geneeskunde)
Aanvullende opmerkingen	<ul style="list-style-type: none"> • Kans is groot dat je veel artikelen met hetzelfde onderwerp krijgt • Liever meer diversiteit • Graag nieuws dat weinig in de headlines voorkomt 	<ul style="list-style-type: none"> • Geen

Google News wordt door mijn twee testpersonen erg positief beoordeeld in de evaluatie. De overzichtelijkheid en gebruiksvriendelijkheid worden als goed ervaren (door één persoon de overzichtelijkheid zelfs als erg goed). Dit is voor een persoonlijke nieuwssite erg belangrijk. Je kan nog zo goed proberen nieuws te personaliseren, als er geen goede lay-out en interactie is met de gebruiker bereik je nog niks.

Over het nieuwsaanbod waren mijn testpersonen ook unaniem: deze vond één persoon goed,

3 Case Studie

de ander zelfs erg goed. Het nieuwsaanbod bij Google News is dus ruim genoeg en, afgaande op vraag vier, komt het nieuws dat aangeboden wordt ook van betrouwbare bronnen.

Op de vraag of de testpersonen dachten dat het nieuws wat ze te zien kregen gerelateerd was aan het nieuws wat ze voorheen hadden gelezen, antwoordde één testpersoon met een volmondig ja. De ander vond het moment afhankelijk. Dit betekent dus wel dat Google Nieuws kijkt naar artikelen waar in het verleden op is geklikt.

Beide testpersonen gaan Google News vaker gebruiken als nieuwssite. De voornaamste reden is vanwege het gebruiksgemak, maar ook snelle koppeling naar Google Mail is hier een reden voor: *“Google news biedt een breed scala aan nieuwsbronnen en ze zijn zeer overzichtelijk geordend. Zeer handig want het leest net zo makkelijk als een krant, je scant snel door de artikelen heen en klikt even als je iets interessant vindt”* (RJ).

“Omdat de site overzichtelijk is, en het ook direct mijn email-account bevat wat wel zo handig is. Ook omdat ze een heel kopje hebben over de gezondheidszorg, iets wat me vanuit mijn studie natuurlijk wel aanspreekt” (JJ).

Zoals al aangegeven in de paragraaf hierboven, was het enige nadeel dat als je op een bepaald onderwerp klikt, de kans groot is dat er erg veel hetzelfde nieuws wordt aanbevolen. Je zult de site dus langer moeten gebruiken (langer dan de vier weken van mijn case studie) om de site een goede indicatie te geven van je voorkeuren en interesses wat betreft nieuws.

3.3.2 Spotback News

Tabel 3: Resultaten Enquête

Vraag	Antwoord CR	Antwoord PB
Overzichtelijkheid	Redelijk	Redelijk
Gebruiksvriendelijkheid	Redelijk	Goed
Nieuwsaanbod	Goed	Redelijk
Betrouwbaarheid Bronnen	Neutraal	Betrouwbaar
Relatie nieuws aan eerder bekeken nieuws	Ja	Soms wel, soms niet
Ga je de site vaker gebruiken?	Nee	Nee
Waarom vaker gebruiken?	<ul style="list-style-type: none"> • Nieuws werd snel eenzijdig • Aanbod politiek nieuws was slecht 	<ul style="list-style-type: none"> • Nieuws niet interessant voor mensen uit Europa • Te Amerikaans
Aanvullende opmerkingen	<ul style="list-style-type: none"> • Site is langzaam en traag 	<ul style="list-style-type: none"> • Als je twee artikelen over een onderwerp positief rate, wordt je gelijk met dat onderwerp overspoelt • Liever een gezonde nieuws-mix

Zoals te zien is aan de antwoorden, wordt Spotback News minder goed beoordeeld dan Google News. De overzichtelijkheid en gebruiksvriendelijkheid van de site worden redelijk bevonden, al ziet PB de gebruiksvriendelijkheid wel als goed. Een redelijke beoordeling van deze twee elementen is niet erg, dit betekent in ieder geval niet dat het erg slecht is.

Het nieuwsaanbod wordt door de één als goed gezien, door de ander als redelijk. De betrouwbaarheid heeft eenzelfde beoordeling: de een ziet het als neutraal (niet goed, niet slecht) de ander als betrouwbaar. Ook hier kun je zeggen dat Spotback News er niet uit springt, maar het ook niet echt slecht doet.

3 Case Studie

De relatie aan eerder bekeken nieuws wordt hetzelfde beoordeeld als Google News, de één vindt er een relatie zitten, de ander af en toe. Op de vraag of de testpersonen de site vaker gaan gebruiken, is het antwoord in beide gevallen 'nee'. De redenen hiervoor zijn duidelijk: het nieuws is te eenzijdig en te Amerikaans. Daarnaast wordt ook aangegeven dat de site vaak traag en langzaam is. Dit speelt natuurlijk ook mee, de gebruikers van tegenwoordig hebben geen zin om lang op hun nieuws te moeten wachten.

3.3.3 Winnaar?

Op basis van deze evaluatie is moeilijk te zeggen welke van de twee sites 'beter' personaliseert. Bij Spotback wordt wel bij beide testpersonen aangegeven dat het nieuws eenzijdig werd. Hier zie je dus eigenlijk dat de personalisatie te sterk werkt. Als je alleen naar persoonlijk nieuwsaanbod kijkt, zou Spotback News dus beter zijn werk doen. Maar ook de site zelf is belangrijk en hier is Google News duidelijk de winnaar. Beide gebruikers zouden de site vaker gebruiken als nieuwssite en vonden de site overzichtelijk.

Je kan stellen dat gebruikers meer behoefte hebben aan zelf hun weg zoeken door de nieuwsberichten heen. Personalisatie is hierbij een leuk extraatje, dat af en toe gebruikt kan worden. Als de hele site zich personaliseert, komt dit het nieuws lezen in ieder geval niet ten goede. Overzichtelijkheid en gebruiksvriendelijkheid worden dus verkozen boven personaliseerbaarheid.

4 Conclusie

Voor het personaliseren van nieuws heb ik twee methoden uitgebreid toegelicht: kansberekening en ontologieën/semantische netwerken. Dit waren de twee technieken die in de literatuur gebruikt en beschreven werden. Op basis van mijn bevindingen in de literatuur heb ik twee persoonlijke nieuwssites, Spotback News en Google News, geanalyseerd en beschreven. Om deze twee websites te testen, heb ik vier mensen per twee personen een site laten gebruiken. De resultaten hiervan waren niet zoals ik ze verwachtte: er was geen duidelijke personalisatie te zien in de resultaten. Uit de evaluatie bleek wel dat qua personalisatie Spotback beter zijn werk deed, maar dat als site Google News beter werkt.

Welke methoden worden er in de literatuur genoemd om voor gepersonaliseerde nieuwssites de voorkeuren van gebruikers te leren kennen, van welke van deze methoden maken vooraanstaande gepersonaliseerde nieuwssites gebruik en hoe goed werken deze sites in de praktijk? Dit is de vraag die ik met mijn onderzoek heb willen beantwoorden. In de literatuur heb ik dus twee methoden gevonden die gebruikt werden: methoden op basis van kansberekening en methoden op basis van ontologieën/semantische netwerken. Daarnaast gebruikt ieder systeem zijn eigen methoden om een gebruikers profiel te genereren, deze te onderhouden en om op basis van dit profiel nieuwsartikelen te filteren. Het onderhoud heeft drie technieken die veel gebruikt worden: directe, indirect en half directe feedback. Het filteren kan op basis van collaborative- of contentbased filtering. Een combinatie hiervan is ook mogelijk.

De vooraanstaande persoonlijke nieuwssites die ik en mijn testpersonen hebben bekeken zijn Google News en Spotback News. De eerste gebruikt impliciete feedback om voorkeuren te leren kennen en het profiel te updaten. Om te filteren gebruikt het content-based filtering. Google News gebruikt, net als vele andere nieuwssites, kansberekening als methoden om waardes van interesses in het gebruikersprofiel te berekenen en nieuws te filteren. Spotback News gebruikt in eerste instantie collaborative filtering om een profiel aan te maken. Zij maken een gemiddeld profiel aan op basis van de profielen van andere gebruikers. Hierna gebruiken zij expliciete feedback om de voorkeuren te leren (gebruikers moeten artikelen raten). Ook Spotback News gebruikt als algemene methode kansberekening.

Ik wilde door mijn case studie bepalen hoe goed beide sites werken als persoonlijke nieuwssites. Op basis van de resultaten van alleen de case studie, lijkt het alsof de sites niet goed werken. Er is geen stijgende trend te zien in het aantal interessante artikelen. Maar op basis van de evaluatie is te zeggen dat ze allebei doen wat ze moeten doen: naar onderwerpen waar eerder op is geklikt zoeken en gerelateerde artikelen presenteren. Helaas gaat dit snel ten kostte van de variëteit van het nieuwsaanbod. Er is dus te zeggen dat er een langere testperiode nodig is om te bekijken of het nieuws ook gevarieerd gepersonaliseerd kan worden.

Op basis van mijn onderzoek moet ik stellen dat, hoewel personalisatie in de theorie iets heel moois lijkt, het in de praktijk vaak betekent dat je teveel in de bepaalde onderwerpen blijft hangen. De nieuwslezers zijn toch geïnteresseerd in een variëteit aan nieuws. Hierdoor hebben zij meer behoefte aan een goed overzichtelijke site als Google News, waarop zij zelf hun artikelen kiezen en er een mogelijkheid tot personalisatie is, dan dat de hele website zich met je nieuwsleesgedrag mee verandert, zoals bij Spotback news.

Referenties

- [ACT00] Liliana Ardissono, Luca Console, and Ilaria Torre. On the application of personalization techniques to news servers on the www. *Lecture Notes of Artificial Intelligence*, 1792:261–272, 2000.
- [ACT01] Liliana Ardissono, Luca Console, and Ilaria Torre. An adaptive system for the personalized access to news. *AI Communications*, 14(3):129–147, 2001.
- [BKA98] Krishna Bharat, Tomonari Kamba, and Michael Albers. Personalized, interactive news on the web. *Multimedia Systems*, 6(5):349–358, Sep 1998.
- [BKB⁺06] E. Banos, I. Katakis, N. Bassiliades, G. Tsoumakas, and I. Vlahavas. *PersoNews: A Personalized News Reader Enhanced by Machine Learning and Semantic Filtering*, volume 4275 of *Lecture Notes in Computer Science*. Springer Berlin / Heidelberg, 2006.
- [Blo09] Spotback Blog. <http://spotback.info/archives/2006/05/>, 11-04-2009.
- [BS97] Marko Balabanovic and Yaov Shoham. Content-based, collaborative recommendation. *COMMUNICATIONS OF THE ACM*, 40(3):66 – 72, March 1997.
- [CCGJ04] Ricardo Carreira, Jaime M. Crato, Daniel Gonçalves, and Joaquim A Jorge. Evaluating adaptive user profiles for news classification. *International Conference on Intelligent User Interfaces*, pages 206 – 212, 2004.
- [COT06] Owen Conlan, Ian O’Keeffe, and Shane Tallon. *Combining Adaptive Hypermedia Techniques and Ontology Reasoning to Produce Dynamic Personalized News Services*, volume 4018 of *Lecture Notes in Computer Science*. Springer Berlin / Heidelberg, 2006.
- [DDG07] Abhinandan Das, Mayur Datar, and Ashutosh Garg. Google news personalization: scalable online collaborative filtering. *Proceedings of the 16th international conference on World Wide Web*, pages 271 – 280, 2007.
- [LLK03] Hung-Jen Lai, Ting-Peng Liang, and Y.C. Ku. Customized internet news services based on customer profiles. *ACM International Conference Proceeding Series*, 50:225 – 229, 2003.
- [LvdG05] P. Lucas and L. van der Gaag. Principles of intelligent systems: a knowledge-based approach. Dictaat Radboud Universiteit, 2005.
- [LYCK08] Ting-Peng Liang, Yung-Fang Yang, Deng-Neng Chen, and Yi-Cheng Ku. A semantic-expansion approach to personalized. *Decision Support Systems*, 45(3):401–412, Jun 2008.
- [MGB⁺04] Mark Maybury, Warren Greiff, Stanley Boykin, Jay Ponte, Chad McHenry, and Lisa Ferro. Personalcasting: Tailored broadcast news. *User Modeling and User-Adapted Interaction*, 14(1):119–144, Feb 2004.

- [MLM03] M. Montaner, B. Lopez, and L.D.A. Mosa. A taxonomy of recommender agents and the internet. *Artificial Intelligence Review*, 19(4):285–330, Jun 2003.
- [New09a] Google News. <http://news.google.com>, 10-04-2009.
- [New09b] Spotback News. <http://news.spotback.com>, 11-04-2009.
- [SK97] Hidekazu Sakagami and Tomonari Kamba. Learning personal preferences on online newspaper articles from user behaviors. *Computer Networks and ISDN Systems*, 29(8-13):1447–1455, Sep 1997.