Probabilistische Gegevensbanken — een korte inleiding

Maarten Fokkinga
DB groep, afd Informatica, fac EWI, Universiteit Twente

Versie van 8 november 2005

Inleiding. Gegevensbanken kennen we allemaal: zonder gegevensbanken zou de wereld niet kunnen functioneren zoals ze nu doet. (O, wat geeft me dat een heerlijk gevoel van onmisbaarheid...) Bijvoorbeeld, de UT heeft gegevensbanken nodig om de studenten te administreren. Vliegmaatschappijen gebruiken gegevensbanken om plaatsboekingen te registreren. De bibliotheek gebruikt gegevensbanken om het boekenbestand vast te leggen, en de uitleningen te registreren. En zo voorts, en zo voorts. Al deze voorbeelden hebben één ding gemeen: er wordt van uit gegaan dat iedere waarde die vast gelegd moet worden, of als antwoord opgeleverd moet worden, geheel bekend is — of helemaal niet (in geval van *Null*). Maar in de wereld om ons heen is er veel onzekerheid, dubbelzinnigheid, onnauwkeurigheid, vaagheid, en zo voorts, tezamen *onbepaaldheid* genoemd. De vraag komt dus op hoe we daar mee om kunnen gaan. Ons uitgangspunt is dat gegevensbeheer niet aan ad-hoc toepassingsprogrammatuur overgelaten moet worden, maar beter door special purpose databasesystemen gedaan kan worden; en dat geldt ook als de gegevens in meer of mindere mate onbepaald kunnen zijn.

Er zijn vele voorstellen gedaan voor het omgaan met onbepaaldheid in gegevensbanken. Eén ervan is het idee van 'probabilistische gegevensbanken'. Dat idee willen we hier duidelijk maken. Dat doen we door eerst van de klassieke gegevensbanken een voorbeeld te geven, dan aan de hand van dat zelfde voorbeeld probabilistische gegevensbanken te definiëren, en tenslotte wat toepassingen (het nut ervan) te laten zien.

Klassieke gegevensbanken. Een gegevensbank is een verzameling feiten; een feit wordt gerepresenteerd als een rij in een tabel. Het antwoord op een vraag aan de gegevensbank is, weer, een verzameling feiten. In het klassieke geval dat alle waarden volkomen bepaald zijn, is een feit *wel* of *niet* waar; een tussenweg is niet mogelijk. Hier is een voorbeeld van een gegevensbank; zeven feiten verdeeld over twee tabellen:

	Naam	Adres	Hobby
$f1$ =	(Apers,	Hengelo,	postzegels)
$f2$ =	(Jonker,	Verwegistan,	suikerzakjes)
$f3$ =	(Apers,	Hengelo,	munten)
$f4$ =	(Peter,	Enschede,	volleybal)

	Groep	Naam
$f5$ =	(DB,	Apers)
$f6$ =	(DB,	Jonker)
$f7$ =	(ES,	Hartel)

Door deze gegevensbank wordt één wereld gedefinieerd; in die wereld zijn precies alle feiten van de gegevensbank waar (en geen andere!):

wereld		feitenverzameling						
w	=	$\{f1,$	$f2,$	$f3,$	$f4,$	$f5,$	$f6,$	$f7\}$

In deze wereld, w, worden vragen als volgt beantwoord:

"Welke hobby's hebben DB-leden?"	postzegels, suikerszakjes, munten
"Waar wonen de DB-leden?"	Hengelo, Verwegistan
"Verzamelt Apers munten?"	ja

Probabilistische gegevensbanken. Bij probabilistische gegevensbanken is een feit met zekere *waarschijnlijkheid* waar (+) of niet waar (−); de + en − waarschijnlijkheden zijn samen 1.0. Hier is een voorbeeld van zo'n probabilistische gegevensbank; er staan dezelfde feiten in als in de vorige, maar nu voorzien van een waarschijnlijkheid:

	Naam	Adres	Hobby	+	−
$f1$ =	(Apers,	Hengelo,	postzegels)	0.6	0.4
$f2$ =	(Jonker,	Verwegistan,	suikerzakjes)	1.0	0.0
$f3$ =	(Apers,	Hengelo,	munten)	0.2	0.8
$f4$ =	(Peter,	Enschede,	volleybal)	0.7	0.3

	Groep	Naam	+	−
$f5$ =	(DB,	Apers)	0.9	0.1
$f6$ =	(DB,	Jonker)	1.0	0.0
$f7$ =	(ES,	Hartel)	1.0	0.0

Van sommige feiten is de +-waarschijnlijkheid op 1.0 gesteld; wanneer van alle feiten de +-waarschijnlijk 1.0 is, hebben we de klassieke gegevensbank weer terug. (Alleen de +-waarden hoeven opgeslagen te worden; de −-waarden kunnen daaruit berekend worden. Voor de eindgebruiker zijn de waarschijnlijkheidskolommen niet zichtbaar.)

Semantiek. We definiëren een *wereld* weer als een verzameling feiten, want voor dergelijke werelden weten we hoe we een vraag er over kunnen beantwoorden. De aanpak is nu dat een probabilistische gegevensbank niet één maar een *stel* werelden definieert, ieder met een zekere mate van waarschijnlijkheid. Bijvoorbeeld, een wereld kan het feit $f1$ = (Apers, Hengelo, postzegels) bevatten, met een waarschijnlijkheid van 0.6, of kan dat feit niet bevatten, met een waarschijnlijkheid van 0.4. Aangezien er in ons voorbeeld 7 feiten zijn, die elk *wel* of *niet* in een wereld als wáár kunnen worden aangenomen, zijn er $2^7 = 128$ verschillende werelden. Eén van de mogelijke werelden is de wereld waarin feiten $f1$, $f2$, $f6$, $f7$ *wel* en $f3$, $f4$, $f5$ *niet* gegeven zijn, wereld $w099$ in Figuur 1 verder op. Voor iedere wereld is de waarschijnlijkheid ervan gelijk aan het product van de +-waarschijnlijkheden van de feiten die er wel in zitten en de −-waarschijnlijkheden van de feiten die er niet in zitten. (Door simpelweg het product te nemen, gaan we er van uit dat de feiten onafhankelijk zijn!) De zojuist genoemde wereld heeft waarschijnlijkheid $f1^+ \cdot f2^+ \cdot f3^- \cdot f4^- \cdot f5^- \cdot f6^+ \cdot f7^+ = 0.6 \cdot 1.0 \cdot 0.8 \cdot 0.3 \cdot 0.1 \cdot 1.0 \cdot 1.0 = 0.0144$. Figuur 1 toont een opsomming van alle werelden met waarschijnlijkheid $\neq 0$ (dus $f2$, $f6$ en $f7$ er wel in), zestien werelden in totaal.

wereld	feitenverzameling							waarschijnlijkheid
$w098 = \{$		$f2,$				$f6,$	$f7\}$	$0.4 \cdot 1.0 \cdot 0.8 \cdot 0.3 \cdot 0.1 \cdot 1.0 \cdot 1.0 = 0.0096$
$w099 = \{f1,$		$f2,$				$f6,$	$f7\}$	$0.6 \cdot 1.0 \cdot 0.8 \cdot 0.3 \cdot 0.1 \cdot 1.0 \cdot 1.0 = 0.0144$
$w102 = \{$		$f2,$	$f3,$			$f6,$	$f7\}$	$0.4 \cdot 1.0 \cdot 0.2 \cdot 0.3 \cdot 0.1 \cdot 1.0 \cdot 1.0 = 0.0024$
$w103 = \{f1,$		$f2,$	$f3,$			$f6,$	$f7\}$	$0.6 \cdot 1.0 \cdot 0.2 \cdot 0.3 \cdot 0.1 \cdot 1.0 \cdot 1.0 = 0.0036$
$w106 = \{$		$f2,$		$f4,$		$f6,$	$f7\}$	$0.4 \cdot 1.0 \cdot 0.8 \cdot 0.7 \cdot 0.1 \cdot 1.0 \cdot 1.0 = 0.0224$
$w107 = \{f1,$		$f2,$		$f4,$		$f6,$	$f7\}$	$0.6 \cdot 1.0 \cdot 0.8 \cdot 0.7 \cdot 0.1 \cdot 1.0 \cdot 1.0 = 0.0336$
$w110 = \{$		$f2,$	$f3,$	$f4,$		$f6,$	$f7\}$	$0.4 \cdot 1.0 \cdot 0.2 \cdot 0.7 \cdot 0.1 \cdot 1.0 \cdot 1.0 = 0.0056$
$w111 = \{f1,$		$f2,$	$f3,$	$f4,$		$f6,$	$f7\}$	$0.6 \cdot 1.0 \cdot 0.2 \cdot 0.7 \cdot 0.1 \cdot 1.0 \cdot 1.0 = 0.0084$
$w114 = \{$		$f2,$			$f5,$	$f6,$	$f7\}$	$0.4 \cdot 1.0 \cdot 0.8 \cdot 0.3 \cdot 0.9 \cdot 1.0 \cdot 1.0 = 0.0864$
$w115 = \{f1,$		$f2,$			$f5,$	$f6,$	$f7\}$	$0.6 \cdot 1.0 \cdot 0.8 \cdot 0.3 \cdot 0.9 \cdot 1.0 \cdot 1.0 = 0.1296$
$w118 = \{$		$f2,$	$f3,$		$f5,$	$f6,$	$f7\}$	$0.4 \cdot 1.0 \cdot 0.2 \cdot 0.3 \cdot 0.9 \cdot 1.0 \cdot 1.0 = 0.0216$
$w119 = \{f1,$		$f2,$	$f3,$		$f5,$	$f6,$	$f7\}$	$0.6 \cdot 1.0 \cdot 0.2 \cdot 0.3 \cdot 0.9 \cdot 1.0 \cdot 1.0 = 0.0324$
$w122 = \{$		$f2,$		$f4,$	$f5,$	$f6,$	$f7\}$	$0.4 \cdot 1.0 \cdot 0.8 \cdot 0.7 \cdot 0.9 \cdot 1.0 \cdot 1.0 = 0.2016$
$w123 = \{f1,$		$f2,$		$f4,$	$f5,$	$f6,$	$f7\}$	$0.6 \cdot 1.0 \cdot 0.8 \cdot 0.7 \cdot 0.9 \cdot 1.0 \cdot 1.0 = 0.3024$
$w126 = \{$		$f2,$	$f3,$	$f4,$	$f5,$	$f6,$	$f7\}$	$0.4 \cdot 1.0 \cdot 0.2 \cdot 0.7 \cdot 0.9 \cdot 1.0 \cdot 1.0 = 0.0504$
$w127 = \{f1,$		$f2,$	$f3,$	$f4,$	$f5,$	$f6,$	$f7\}$	$0.6 \cdot 1.0 \cdot 0.2 \cdot 0.7 \cdot 0.9 \cdot 1.0 \cdot 1.0 = 0.0756$
								1.0000

Figuur 1: *De werelden gedefinieerd door de probabilistische gegevensbank. Alleen de werelden met positieve waarschijnlijkheid zijn hier opgesomd.*

Om de semantiekdefinitie te voltooien moeten we ook nog aangeven hoe vragen aan de gegevensbank worden beantwoord. Dat doen we aan de hand van deze voorbeeldvraag: "Welke hobby's hebben DB-leden?". We beantwoorden deze vraag door *in iedere mogelijke wereld* de vraag te beantwoorden; bedenk dat een wereld volkomen bepaald is, en dus geen principiële moeilijkheden geeft in de beantwoording:

In wereld ↓,	met waarschijnlijkheid ↓,	is het antwoord ↓.
w098 $= \{f2, f6, f7\}$	0.0096	suikerzakjes
w099 $= \{f1, f2, f6, f7\}$	0.0144	suikerzakjes
w102 $= \{f2, f3, f6, f7\}$	0.0024	suikerzakjes
w103 $= \{f1, f2, f3, f6, f7\}$	0.0036	suikerzakjes
w106 $= \{f2, f4, f6, f7\}$	0.0224	suikerzakjes
w107 $= \{f1, f2, f4, f6, f7\}$	0.0336	suikerzakjes
w110 $= \{f2, f3, f4, f6, f7\}$	0.0056	suikerzakjes
w111 $= \{f1, f2, f3, f4, f6, f7\}$	0.0084	suikerzakjes
w114 $= \{f2, f5, f6, f7\}$	0.0864	suikerzakjes
w115 $= \{f1, f2, f5, f6, f7\}$	0.1296	postzegels, suikerzakjes
w118 $= \{f2, f3, f5, f6, f7\}$	0.0216	suikerzakjes, munten
w119 $= \{f1, f2, f3, f5, f6, f7\}$	0.0324	postzegels, suikerzakjes, munten
w122 $= \{f2, f4, f5, f6, f7\}$	0.2016	suikerzakjes
w123 $= \{f1, f2, f4, f5, f6, f7\}$	0.3024	postzegels, suikerzakjes
w126 $= \{f2, f3, f4, f5, f6, f7\}$	0.0504	suikerzakjes, munten
w127 $= \{f1, f2, f3, f4, f5, f6, f7\}$	0.0756	postzegels, suikerzakjes, munten
	1.0000	

Sommeren we voor ieder antwoord de waarschijnlijkheden van de werelden waarin dat antwoord opgeleverd wordt, dan krijgen we de volgende lijst:

postzegels, suikerzakjes	0.4320	$(= 0.1296 + 0.3024)$
suikerzakjes	0.3880	$(= 0.0096 + \cdots + 0.0864 + 0.2016)$
postzegels, suikerzakjes, munten	0.1080	$(= 0.0324 + 0.0756)$
suikerzakjes, munten	0.0720	$(= 0.0216 + 0.0504)$
	1.0000	

Deze lijst is het beste antwoord dat gegeven kan worden, maar heeft wel één groot nadeel: zo'n lijst kan 2^n veel regels bevatten, wanneer er n items in een antwoord kunnen voorkomen. Dat is veel te veel voor practisch gebruik. Practischer is het om voor ieder van de n items aan te geven met welke waarschijnlijkheid het in een antwoord voorkomt. Dat levert de volgende lijst (gerangschikt naar afnemende waarschijnlijkheid):

suikerzakjes	1.0000	$(= 0.4320 + 0.3880 + 0.1080 + 0.0720)$
postzegels	0.5400	$(= 0.4320 + 0.1080)$
munten	0.1800	$(= 0.1080 + 0.0720)$

Deze uitkomst geeft aan dat met zekerheid suikerzakjes gespaard worden door DB-leden, dat postzegels gespaard worden met waarschijnlijkheid 0.5400, en munten met waarschijnlijkheid 0.1800. (Opgeteld zijn deze waarschijnlijkheden uiteraard meer dan 1.000.)

Nog meer voorbeeldvragen. Net zo als hierboven wordt de vraag "Waar wonen de DB-leden?" beantwoord met:

Hengelo, Verwegistan	0.6120	$(= 0.0096 + \cdots + 0.0864 + 0.2016)$
Verwegistan	0.3880	$(= 0.1296 + 0.0216 + 0.0324 + 0.3024 + 0.0504 + 0.0756)$
	1.0000	

Op de vraag "Verzamelt Apers munten?" is het antwoord is 'ja' in de acht werelden waarin $f3$ zit, met gezamenlijke waarschijnlijkheid 0.2000, en 'nee' in de acht werelden waarin $f3$ niet zit, met gezamenlijke waarschijnlijkheid 0.8000:

nee	0.8000
ja	0.2000
	1.0000

Deze uitkomst vind je ook direct uit de eerste van de twee gegeven tabellen van de probabilistische gegevensbank; want de tweede tabel zegt niets over hobby's.

Toepassingen. Er zijn verscheidene manieren waarop de waarschijnlijkheden in de tabellen en antwoorden tot stand komen en uitgebuit kunnen worden:

- *Dubbelzinnigheid, tegenstrijdigheid.* Wanneer twee gegevensbronnen worden samengevoegd (je synchroniseert het elektronische adresboek op je PDA met die van je PC), kan één bron aangeven dat Apers postzegels verzamelt terwijl de andere bron zegt dat Apers munten verzamelt. Zelfs in geval dat personen hooguit één hobby hebben, worden beide feiten opgenomen, maar ieder met een waarschijnlijkheid van 0.5. (Zo doende kunnen we met *tegenstrijdige* gegevens omgaan.) Als er méér dan twee bronnen meedoen, kunnen de waarschijnlijkheden van 0.5 afwijken, of als één van de bronnen om wat voor reden dan ook voorrang krijgt kan de waarschijnlijkheid van beide feiten ook anders zijn dan half-half.

- Wanneer twee bronnen worden samengevoegd, kan het zijn dat de bedoelde interpretatie van *schema*'s van de bronnen verschillen: bij de één betekent *naam* de voornaam van een persoon terwijl bij de ander de achternaam bedoeld wordt. *Hmmmmmmmmm, hoe leidt dit tot probabilities?*

- *Onnauwkeurigheid.* Wanneer de gegevensbank gevuld wordt met sensorgegevens die de locatie van iemand aanduiden, kan het zijn dat de sensorgegevens niet geheel *nauwkeurig* zijn. Bij een GPS localisering van Apers kan enerzijds Hengelo als locatie gevonden worden en anderzijds Enschede. Beide locaties kunnen in de gegevensbank opgenomen worden, ieder met 0.5 waarschijnlijkheid, of een andere waarschijnlijkheid als daarvoor andere aanwijzingen zijn.

- *Overeenkomstigheid.* Wanneer van Peter Apers het adres gevraagd wordt, kan, bij de voorbeeldgegevensbank, van zowel Peter als ook van Apers het adres gegeven worden, ieder met een waarschijnlijkheid van 0.5, of met grotere waarschijnlijkheid voor het adres van Apers, wanneer bekend is 'Apers' veel onderscheidener is dan 'Peter'.

- *Overeenkomstigheid.* Wanneer van 'Aders' het adres gevraagd wordt, wordt in het klassieke geval niets opgeleverd, terwijl bij probabilistische gegevensbanken de adressen van 'Apers', 'Jonker' en 'Peter' opgeleverd kunnen worden, ieder met een waarschijnlijkheid die berekend wordt uit de mate waarin 'Aders' overeenkomt met de naam (dus een hogere waarschijnlijkheid bij het adres van Apers dan bij het adres van Jonker). Door deze manier kan de gebruiker in veel van de gevallen nog een zinvol antwoord gegeven worden waar in het klassieke geval het antwoord leeg is.

- *Voorkeur.* Wanneer een gebruiker een vraag formuleert, kan dat in het klassieke geval alleen maar een exacte vraag zijn, zonder onbepaaldheid er in. Maar we kunnen nu ook benaderingen toestaan, zoals 'ongeveer 4' en 'ongeveer 500 euro'. De beantwoording van een vraag *begint* dan met het toekennen van 'scores' aan feiten in de database; de scores geven aan hoe goed het feit past bij de vraag. Die scores worden verder precies zo behandeld als de waarschijnlijkheden. Dus het antwoord wordt in een probabilistische gegevensbank uitgerekend, en bestaat weer uit een opsomming van mogelijkheden, ieder voorzien van de score (waarschijnlijkheid) van de mate waarin het antwoord aan de exacte vraag voldoet.

- *Rangschikking.* Bij probabilistische gegevensbanken zullen de antwoorden al gauw grote opsommingen zijn. Niet alleen worden die opgesomd in volgorde van afnemende waarschijnlijkheid, maar ook kun je afspreken dat alleen de top-10 wordt opgeleverd (en op verzoek pas de rest).

Onderzoek. Probabilistische gegevensbanken zijn al lang bekend, maar pas recentelijk echt in de belangstelling komen staan door twee redenen. Ten eerste is de behoefte aan het beheer van enigszins onbepaalde gegevens recentelijk enorm gegroeid door het karakter van toepassingen. Ten tweede zijn er recentelijk theorieën en technieken ontwikkeld die een efficiënte beantwoording van vragen soms, maar nog lang niet altijd, mogelijk maakt. Veel onderzoek is nog nodig om tot beter begrip van toepasbaarheid te komen, en in méér gevallen ook vragen efficiënt te kunnen beantwoorden.

Bronnen. Voor deze uiteenzetting is gebruik gemaakt van het werk van Dan Suciu.