Andrea Scharnhorst, DANS KNAW

'BiG Grid and beyond'

Wikipedia als proefkonijn

De hoeveelheid data op het internet is duizelingwekkend, maar die is alleen bruikbaar dankzij een goede indeling. Wikipedia categoriseert sinds 2004 artikelen. Dat gebeurt zonder regels, volgens de geest van de vrije digitale encyclopedie. Andrea Scharnhorst, leider van de eResearch groep van DANS, onderzocht met haar team hoe er zonder regels toch gestructureerde en hiërarchische classificatiesystemen ontstaan. 'Het was een van mijn leukste projecten ooit, omdat er zoveel is uitgekomen.'

Nieuw

DANS is een instituut van de Koninklijke Nederlandse Academie van Wetenschappen (KNAW) en NWO, gericht op het archiveren en ontsluiten van digitale onderzoeksdata. Met haar team van de Knowledge Space Lab wilde Scharnhorst de categorieën in de Engelse Wikipedia monitoren, maar dit leek al snel vast te lopen. 'We hadden 2.8 TeraByte aan data', vertelt Scharnhorst. 'Maar we hadden geen idee hoe we die moesten downloaden en werden daardoor een beetje wanhopig. Projecten van dit formaat waren nieuw voor ons. Toen we eind 2009 met BiG Grid om de tafel gingen, werden onze zorgen weggenomen en konden we snel aan de slag. De parallelprocessing hebben we op het grid laten lopen.'

Het onderzoek van Scharhorst is gericht op het bouwen van nieuwe interfaces, zodat users die data beter kunnen benutten en vinden. 'Ik wilde weten hoe het categoriesysteem van de Engelse Wikipedia zich ontwikkelt. Over de eeuwen heen hebben mensen geprobeerd kennis te ordenen en in systemen te stoppen. De Belgische bibliograaf Paul Otlet hield zich eind negentiende eeuw al bezig met een systeem om alle kennis ter wereld toegankelijk te maken en bedacht de Universele Decimale Classificatie (UDC). Dit classificatiesysteem wordt nog steeds gebruikt in bibliotheken, archieven en musea. Die oude en soms stroeve systemen wilde ik meenemen in dit onderzoek en daarom hebben we Wikipedia vergeleken met de UDC.'

Kennis ordenen

Door Wikipedia als proefkonijn te gebruiken, wilde Scharnhorst informatie verkrijgen om de kwaliteit van informatievoorziening te verbeteren. 'In Wikipedia mag iedereen categorieën toevoegen. Dat zijn losse tags en nergens staat beschreven waaraan die moeten voldoen. Vervolgens discussiëren de vrijwillige Wikipedia redacteuren over de plaats die de tag krijgt binnen de al bestaande categorieën. Uiteindelijk heb je categoriepagina's die ook onder elkaar in relatie staan en vormt zich een ordeningssysteem. We wilden die categorieën inhoudelijk bekijken.'

Evolutie

Scharnhorst ontdekte dat in het categoriesysteem van Wikipedia veel wordt gewijzigd. 'In de huidige Wikipedia heb je een categoriepagina die heet Main Topic Classifications. Als je de geschiedenis van deze top categorie bekijkt, kom je tot 2006. Daarvoor had die categorie een andere naam. Ook de links blijken heel snel te veranderen.

Toen we het Wikipedia project met BiG Grid begonnen, vonden we wel een aantal studies over Wikipedia categoriesystemen. Maar niemand had naar de evolutie van die categoriesystemen van Wikipedia gekeken. Wij zijn de eerste die hier naar keken, en dat zou niet mogelijk zijn geweest zonder BiG Grid. Het was een van mijn leukste projecten ooit. Het heeft tot wetenschappelijke publicaties geleid en we hebben inzicht gekregen in hoe we de ontknoping en ontsluiting van databestanden wereldwijd mogelijk kunnen maken. Ook is er een en een prachtige poster gemaakt die de vergelijking van Wikipedia met UDC visualiseert.'

Klik hier voor referenties