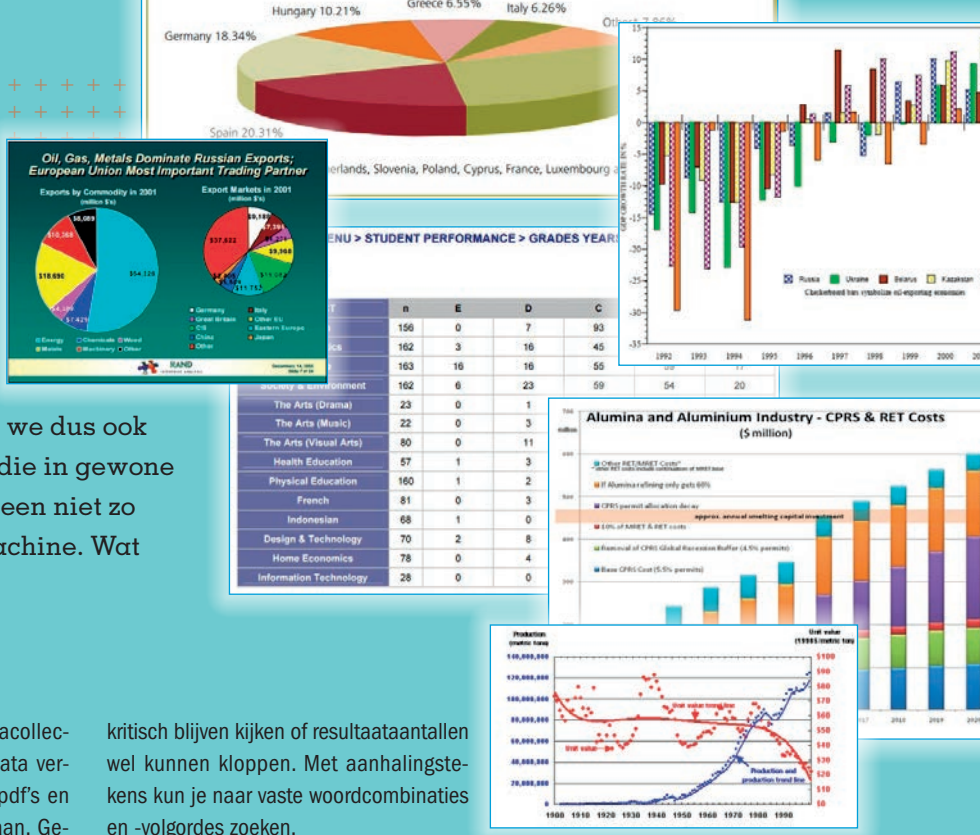
DAT ZOEKEN WE OP

Zanran



Data worden als steeds belangrijker informatiebron onderkend. Daar willen we dus ook gericht naar kunnen zoeken. Voor data die in gewone webdocumenten voorkomen, is Zanran een niet zo bekende maar wel interessante zoekmachine. Wat kan daarmee wel en niet?

Door: Eric Sieverts

Behalve in gespecialiseerde datacollecties zitten ook ontzettend veel data verborgen in gewone webpagina's, pdf's en spreadsheets die op internet staan. Gewone zoekmachines zoeken daar wel in, maar bieden geen filtermogelijkheden om daarin 'data' te herkennen. Zanran doet dat wel. Die pikt juist grafieken, schema's, tabellen, staafdiagrammen en dergelijke uit webpagina's en pdf's. En uiteraard Excel-sheets die haast per definitie 'data' bevatten.

Opmerkelijke eigenschappen van Zanran:

> in resultatenlijsten krijg je met muisover meteen pop-ups te zien van de tabel of grafiek waarop het zoekresultaat gebaseerd is, wat selecteren vereenvoudigt;

> bij aanklikken van een pdf-resultaat wordt daarin meteen doorgescrolld naar de plek waar de betreffende tabel of grafiek staat, ook al is dat pas op bladzijde 30.

Klassiek Booleaans

In Zanran kun je Booleaans combineren met AND, OR en NOT. Bij gemengde AND/OR-opdrachten moeten ge-OR-de termen klassiek tussen haakjes staan. AND mag je in principe weglaten, maar soms gebeuren er gekke dingen als je dat in zo'n gemengde opdracht doet. Je moet dus kritisch blijven kijken of resultaataantallen wel kunnen kloppen. Met aanhalingstekens kun je naar vaste woordcombinaties en -volgordes zoeken.

Varianten en synoniemen

Zanran kent geen truncatie, maar zoekt wel automatisch op woordstammen. Dat gaat verder dan alleen enkel- en meervoud; ook werkwoordsvormen worden meegenomen. Dat is niet altijd een voordeel. Als je op *corn* (mais) zoekt, blijkt ook gezocht te zijn op *corning*, de naam van een bedrijf in een heel andere sector. Bij zoeken op 'exacte phrases' wordt, anders dan bij andere systemen, nog steeds op woordstammen gezocht. Met *'pork export'* vind je dus ook *'pork exports'* en *'pork exporters'*. Erg handig, maar dat maakt het onmogelijk om op echt exacte woorden te zoeken.

Zanran gaat nog een stukje verder. Aan de vette woorden in de zoekresultaten zie je dat in veel gevallen ook automatisch op synoniemen wordt gezocht. *Dutch* geeft ook *Netherlands*, *ca* ook *california*, *corn* ook *maize*, enzovoort. Met OR is makkelijk te controleren dat die synoniemen ook echt volledig worden meegenomen.

Filters

Filtermogelijkheden zijn er op landen/domeinen, op recentheid en op filetype. Filteren op Excel-sheets kan nuttig zijn als je gegevens zelf ook meteen weer in een spreadsheet zou willen verwerken. Maar als gewenste gegevens alleen in andere file-formaten blijken voor te komen, zul je het daar natuurlijk mee moeten doen.

Niet alles even vers

In mijn zoekcursussen laat ik soms ook in Zanran zoeken. Daardoor heb ik nog oude gegevens over zoekresultaten. Helaas onthullen die een zwakke kant van Zanran. Voor die vragen krijg ik nu nog precies dezelfde aantallen resultaten als een jaar geleden. De index wordt dus niet erg frequent geüpdatet. Hoe zit het dan met dat recentheidsfilter? Inperken op 'laatste 6 maanden' geeft wel resultaat, maar daar blijken ook documenten van 10 jaar geleden bij te zitten. Dat filter blijkt dus niet erg zinnig.

Alternatieven

Zijn er, gezien die beperkingen, nog alternatieven waar je wel recent materiaal vindt? Eigenlijk alleen voor speciale situaties. Als het je om Excel sheets gaat, kun je met gewone Google zoeken op bijvoorbeeld *milk exports filetype:xls OR filetype:xlsx*. (Let op de OR, want Google neemt die twee Excel-versies niet automatisch samen.) Dat geeft soms nog wat meer resultaten dan Zanran met Excelfilter.

Een andere mogelijkheid is de – tamelijk onbekende – experimentele tabellenzoeker van Google: research.google.com/tables. Die zoekt naar tabellen die in webpagina's en pdf's voorkomen, maar dus niet op grafieken of staafdiagrammen. Bovendien levert die altijd veel minder resultaten. Voor *milk export netherlands* zegt hij 3189 resultaten te hebben, maar de laatste die je te zien kunt krijgen is nummer 39. De 802 resultaten van Zanran op die vraag krijg je tenminste echt. <

Eric Sieverts, redacteur van IP en freelance docent en adviseur

URL	zanran.com
Booleaans combineren	ja
Truncatie	nee (wel automatisch woordstammen)
Speciale zoekvelden	nee (wel zoeken op documentsoorten)
Advanced zoekscherm	nee
Parametrische filters	nee (wel filters vooraf of achteraf)
Ook in Google	nvt
Semantische markup	nvt