

Voor de dolende webspeurder

Visualisatie van documenten

Jan Kraak j.kraak@rc.rug.nl

Web-zoekmachines zoals Alta Vista of Google voldoen goed wanneer de relevante trefwoorden van documenten bekend zijn. Maar ze zijn minder geschikt voor het verkennen en het krijgen van een overzicht van een onbekende collectie documenten. Daarin kan verbetering worden gebracht door de documenten te visualiseren, zodat verbanden zichtbaar worden en exploratie mogelijk wordt. Na een inleiding worden in dit artikel drie voorbeelden van document-visualisatie gegeven, met als laatste de WEBSOM-methode gebaseerd op een tweedimensionaal neurale netwerk waarmee reeds 7.000.000 documenten tegelijk kunnen worden gevisualiseerd.

Een bijzondere winkel

In Leeuwarden bevindt zich aan de Westerplantage de bijzondere winkel van de firma Auke Rauwerda, gevestigd in een aantal naast elkaar gelegen panden. Daardoor verwacht je bij binnenkomst in een grote ruimte terecht te komen en **niet** in een kleine, tamelijk donkere, ruimte met een smalle toonbank, waarachter enkele vriendelijke mannen staan, tegen wie je zegt wat je wilt hebben.

Nadat je dat hebt gedaan, snelt een van de mannen weg naar de magazijnen om het verlangde voor je op te halen. Vaak duurt het even, want de magazijnen hebben een enorme omvang. Zonder dat het van buiten opvalt, beslaan ze heel wat huizenblokken van Leeuwarden ? maar nu overdrijf ik waarschijnlijk een beetje. Duidelijk is dat ze niet vaak ?nee? hoeven te verkopen, al krijg je de inhoud van de magazijnen niet te zien. Ze hebben een vaste klantenkring van mensen die precies weten wat ze willen hebben en ook een overzicht in hun hoofd hebben van wat ze bij Rauwerda verkopen. Funshoppers, die zonder een vooropgezet doel een winkel binnen komen en wat rondsnuffelen, zul je er niet aantreffen, want daar leent de winkel zich niet voor.

Deze bijzondere winkel is overigens enige jaren geleden tentoongesteld geweest in het Fries Museum in Leeuwarden.



Huidige web-zoekmachines vertonen gelijkenis met de manier waarop je geholpen wordt in bovenstaande winkel van Auke Rauwerda in Leeuwarden.

Zoekmachines ontoereikend

Waarom zoveel aandacht voor een winkel, die in zeker opzicht niet meer van deze tijd is, zult u zich afvragen? Wel, om duidelijk te maken dat een veel gebruikt hulpmiddel op internet eigenlijk ook niet meer van deze tijd is, aangezien dat ongeveer net zo te werk gaat als het bedienend personeel in de winkel van Rauwerda.

Ik bedoel de zoekmachines op internet zoals Alta Vista of Google waar, dat moet worden toegegeven, een fantastische technologie achter schuilt. Maar evenals bij Rauwerda, moet je *je wensen in de vorm van een aantal trefwoorden formuleren in een ?loketje?, zonder enig zicht op de informatie op het web en de wijze waarop er wordt gezocht.*

Formuleer je je zoekopdracht verkeerd, dan krijg je niets of - wat erger is - een lijst die zo groot is, dat je er in verzuipt. Dan is men bij de firma Rauwerda heel wat gebruikersvriendelijker: daar zeggen ze tenminste meteen wanneer ze iets niet hebben en komen ze niet terug met vrachtwagens vol artikelen als je je vraag verkeerd hebt geformuleerd, maar ze helpen je om je vraag beter te formuleren.

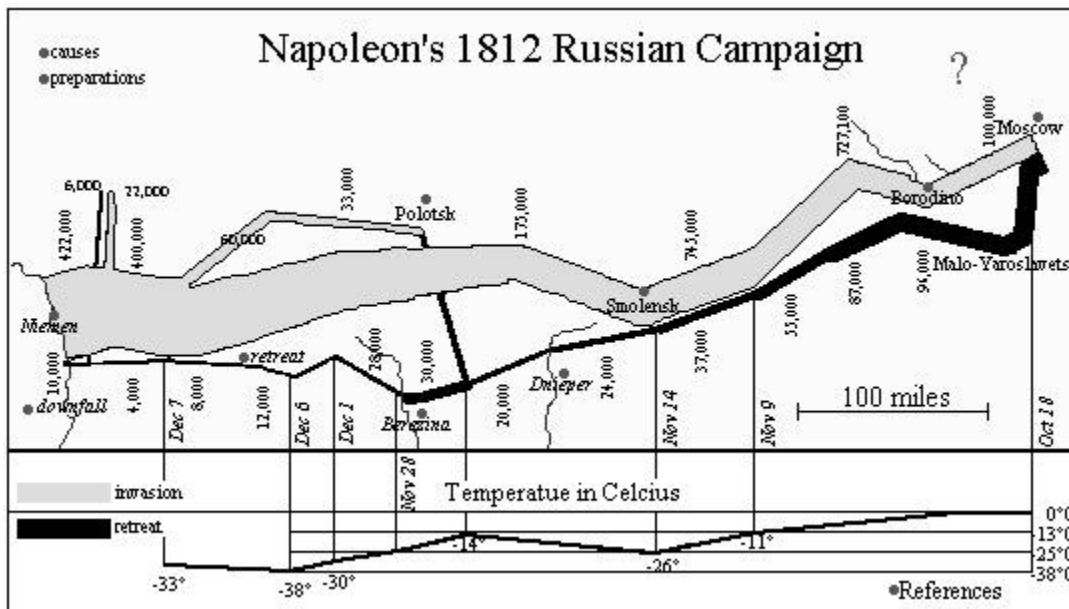
Visueel overzicht gewenst

Met een web-zoekmachine zou je, als aanvulling op het zoeken op trefwoorden, eigenlijk ook de mogelijkheid willen hebben om een overzicht te krijgen over alle webdocumenten betreffende een bepaald onderwerp. Zo'n overzicht zou de vorm kunnen hebben van de inhoudsopgave van een boek, maar een grafische voorstelling, ofwel een visualisatie, van de inhoud van tekstdocumenten is veel beter. Want ons menselijke visuele systeem is bij uitstek het instrument om verbanden in grote hoeveelheden gevisualiseerde data te ontdekken. Het is nu

de kunst om de documenten zodanig te visualiseren, dat ons oog er onmiddellijk verbanden, uitzonderingen etc. in ziet.

Geslaagde visualisaties

Er zijn op allerlei terreinen al heel wat geslaagde visualisaties gemaakt van grote hoeveelheden data, die in 1812 oogopslag iets duidelijk maken. Velen noemen de grafiek van de veldtocht van Napoleon naar Rusland in 1812 door de Franse cartograaf Minard de beste visualisatie aller tijden. Ook Nederlanders namen daar aan deel.



De veldtocht van Napoleon naar Rusland in 1812

Door middel van de bandbreedte is het aantal deelnemende c.q. nog in leven zijnde soldaten op een bepaald traject gevisualiseerd. De bovenste lichte band geeft de intocht weer en de zwarte de terugtocht. Daaronder staan tevens de temperatuur en de data aangegeven. Van de 422.000 soldaten aan het begin haalden slechts ongeveer 10.000 de eindstreep. Bij de overtocht van de rivier de Berezina zakten duizenden soldaten door het ijs.

Omdat visualisatie zijn nut heeft bewezen, ligt het voor de hand dit krachtige middel ook toe te passen op document-visualisatie en op het verkennen van internet. Om grote hoeveelheden data te exploreren en om verbanden erin te vinden zonder dat er vooraf een model aanwezig is.

Document-visualisatie

Als onderdeel van information visualization, een nieuwe visualisatie-toepassing naast bestaande zoals scientific visualization en het tekenen van XY-grafieken, wordt er sinds enige tijd onderzoek gedaan naar de visualisatie van grote hoeveelheden documenten: in het ideale geval honderden miljoenen webdocumenten.

Daarbij moet het niet-visuele worden omgezet in het visuele. De termen die daarbij worden gebruikt zijn *visual data mining*, *text visualization*, *text data mining*, *text mining*, *visual text analysis*, *document visualization*, *web-based information visualization*, *document classification* etc.

Na lang zoeken op het web met Alta Vista en Google, dat behalve veel waardevols ook de nodige ? eendagsvliegen? opleverde, ontdekte ik eindelijk (?als een schat onderin de mijn? ? om de metafoor van de mijn te gebruiken) het goed geschreven proefschrift van Krista



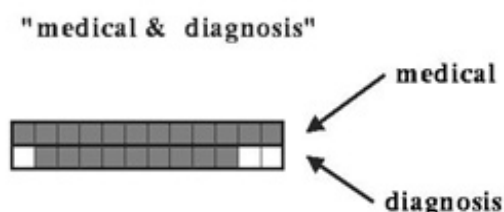
Lagus, getiteld 'Text Mining with WEBSOM' (2000) dat ik kan aanbevelen. Ook het boek 'Information Visualization' van Robert Spence (2001), en dan vooral het laatste hoofdstuk 'Document Visualization', is de moeite waard om te lezen.

Als een introductie tot het nieuwe gebied van tekstvisualisatie behandel ik hierna in het kort een drietal technieken, met als laatste WEBSOM.

Krista Lagus schreef een waardevol proefschrift over text mining

Verdeling van trefwoorden

Met 'TileBars', van Marti A. Hearst van de Universiteit van Californie, Berkeley (1995, 1999), kan de verdeling van trefwoorden in documenten, die bijvoorbeeld zijn gevonden met een zoekmachine, worden gevisualiseerd.



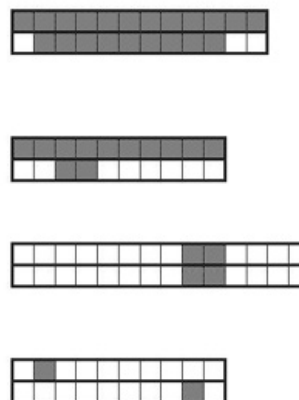
Stel, je hebt documenten op de twee trefwoorden *medical* en *diagnosis* gezocht, en je bent geïnteresseerd in het verband tussen beide begrippen. Elk document wordt dan door twee liggende staven afgebeeld.

De staven zijn onderverdeeld in 'tegels' overeenkomend met tekstsegmenten, zoals paragrafen of alinea's. De grijswaarde van elke tegel is evenredig met het aantal malen dat het trefwoord daarin voor komt. In een witte tegel komt het trefwoord dus niet voor.

Visualisatie d.m.v. een TileBar

Door nu de TileBars voor een aantal documenten onder elkaar te tekenen, ziet u vrijwel in één oogopslag de verdeling van trefwoorden over de documenten. De staaf lengte is tevens evenredig met de documentlengte.

Uit de hiernaast staande visualisatie van vier documenten ziet u dat alleen in het bovenste document in vrijwel alle segmenten beide trefwoorden voorkomen, zodat het aannemelijk is dat het document over het verband tussen beide trefwoorden gaat. De documenten 2 en 3 kennen slechts twee gemeenschappelijk segmenten, en het laatste geen.



Zakken vol woorden

Bij de TileBars moeten de trefwoorden bekend zijn.

Tevens is het aantal te visualiseren documenten beperkt, vanwege de ruimte die ze op een beeldscherm innemen. Als u een nieuw terrein van onderzoek betreedt, kent u de inhoud van de documenten, waaronder relevante trefwoorden, nog onvoldoende. Dan wilt u graag, dat elk document *automatisch* door een aantal parameters wordt samengevat.

Aan de hand van dergelijke parameters moet dan een visualisatie van *het geheel* van documenten worden gemaakt, daarbij dient er plaats te zijn voor een groot aantal documenten. Bij voorkeur moeten de documentparameters *getallen* zijn want daarvoor zijn er reeds allerlei visualisatietechnieken bekend.

Het zal duidelijk zijn dat het vinden van een getalsmatige representatie van documenten geen gemakkelijke opgave is. Alleen bij de verzameling woorden voor kleuren is dat exact te doen. U vertaalt dan namelijk elke kleur in de bijbehorende waarden voor de componenten rood, groen en blauw: rood = (1,0,0) en wit = (1,1,1) etc.

Als eerste benadering vat men bij veel tekst-visualisatiemethoden een document op als een 'zak vol woorden'. Van elk woord telt men de *frequentie*. Men zondert daarbij 'vulwoorden' als 'en', 'dus' etc. uit. Ook worden URL's, images etc. niet meegenomen tijdens het tellen. Zo'n woordfrequentietelling levert per document meestal een groot aantal getallen (N) op: in de orde van 1000 of nog veel meer.

TileBars voor vier documenten

Hoog-dimensionale bloem platdrukken

Per document worden de woordfrequenties opgevat als een punt in een N-dimensionale ruimte, met per dimensie 馃 woord. Alle documenten voor een bepaald onderzoek te samen vormen dan een *puntenwolk* in een *hoog-dimensionale ruimte*, die niet zonder meer gevisualiseerd kan worden. Dat kan alleen voor hooguit drie dimensies, want meer dimensies hebben we in de gewone wereld niet.

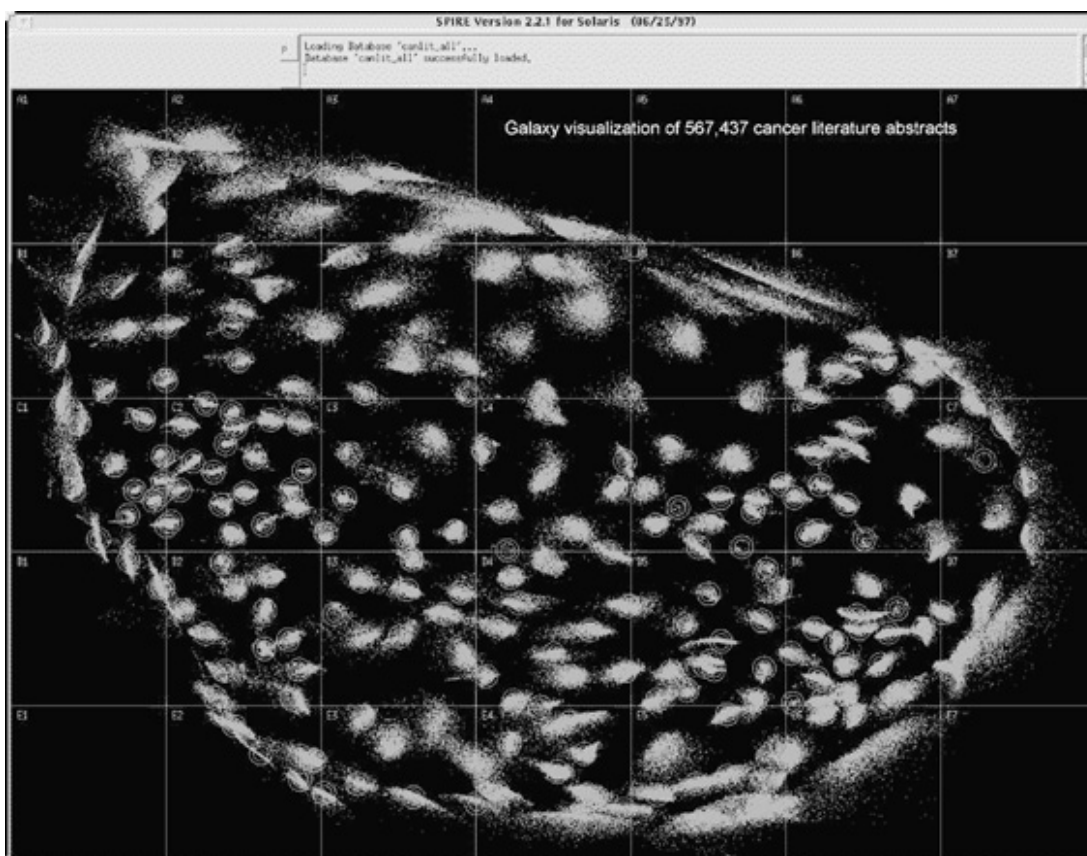
De verschillende document-visualisatiemethoden onderscheiden zich in de wijze waarop de hoog-dimensionale wolk op het tweedimensionale (2D) vlak wordt geprojecteerd, waarbij zo weinig mogelijk karakteristieke eigenschappen verloren mogen gaan. U heeft vast wel eens een bloem tussen de bladzijden van een boek gelegd om te drogen. Daarbij verschoof en plooid u de onderdelen van de bloem zodanig dat bij het platdrukken de bloem goed herkenbaar bleef. In feite is dit een projectie van 3D naar 2D, bij document-visualisatie gaat het evenwel om het platdrukken van een hoog-dimensionale bloem. Projectie op 3D kan desgewenst gebeuren in onze toekomstige CAVE, maar die beschouwen we hier niet.

Sterrenwolken

Bij de 'Galaxies-visualisatie' worden cluster- en principale componenten-analyse toegepast. Het resultaat is een wolk van punten die doet denken aan de nachtelijke sterrenhemel, waarin we als mens gewend zijn om daar patronen in te herkennen. Verwante documenten klonteren samen, terwijl onverwante documenten ver uit elkaar liggen.

Met Galaxies kan men verder de groepstructuur van de documenten onderzoeken, de inhoud van afzonderlijke documenten opvragen en - wat belangrijk is - trends in de *tijd* opsporen. Zo heeft u snel door welk type onderzoek tot niets heeft geleid.

Galaxies is 馃 van de vele hulpmiddelen ontwikkeld door het Pacific Northwest National Laboratory voor visuele tekstanalyse onder de naam 'SPIRE': Spatial Paradigm for Information Retrieval and Exploration.



Sterrenwolk-visualisatie van de samenvattingen van 56.7439 artikelen over kanker

Behalve de sterrenhemel, passen de SPIRE-hulpmiddelen nog andere metaforen toe zoals:

- het reli馃 van een natuurlijk terrein, belangrijke thema's zijn door heuvels weergegeven (ThemeView)

- het licht van sterren (Starlight) - regenbogen voor de weergave en de exploratie van verbanden (Rainbows)
- rivieren om het tijdsverloop van patronen, trends, relaties in grote collecties documenten etc. weer te geven (ThemeRiver).

Het in ontwikkeling zijnde WebTheme past SPIRE toe op webvisualisatie.

SOM/Kohonen maps

Hierbij wordt het Self-Organizing Map (SOM) algoritme uit de theorie van de neurale netwerken toegepast om de hoog-dimensionale puntenwolk in het tweedimensionale vlak te projecteren, en wel zodanig dat verwante documenten dicht bij elkaar te komen liggen.

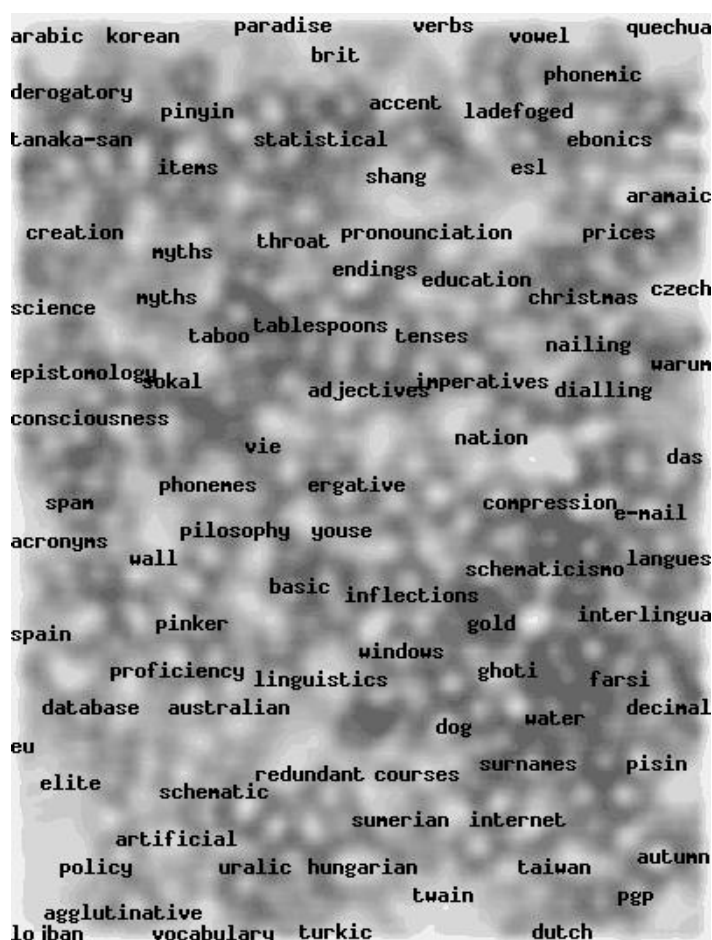
Het SOM- of Kohonen-algoritme, genoemd naar zijn uitvinder Kohonen, is een van de meest gebruikte neurale netwerk algoritmen. Het wordt gebruikt c.q. gerefereerd in meer dan 4000 wetenschappelijke artikelen. Het algoritme staat onder meer beschreven in het reeds genoemde proefschrift van Krista Lagus, werkzaam op het door Teuvo Kohonen opgerichte Neural Network Research Centre van de Technische Universiteit van Helsinki.

Newsgroup ?sci.lang? gevisualiseerd

Lagus geeft een aantal voorbeelden van WEBSOM, zoals de webtoepassing van SOM worden genoemd. Hier tonen we de visualisatiekaart van de 32617 artikelen van newsgroup ?sci.lang? over aspecten van talen, overgenomen van de WEBSOM-website.

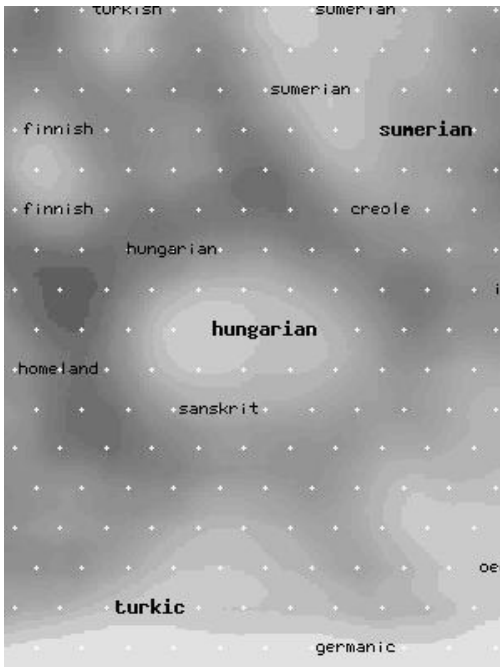
De grijswaarde geeft de dichtheid van documenten aan: des te lichter de kleur des te meer documenten. In de kaart staan trefwoorden voor de artikelen, op de website zijn ze nader verklaard.

Door ?hungarian?, midden onderin de kaart aan te klikken, wordt op een deel van de omgeving van dat punt ingezoomd waarin door middel van witte punten de neuronen staan aangegeven. Door hier weer te klikken krijgt u een artikel uit de newsgroup ?in het echt? te zien.



Kohonen-map van 32.617 artikelen van de newsgroup sci.lang

Ingezoomd op ?hungarian?



Op de WEBSOM-website staan nog drie andere SOM-visualisaties, waaronder 馱 van 1.000.000 artikelen uit 80 verschillende newsgroups.

Parallele implementatie

Door Krista Lagus en haar collega's is uitgebreid met WEBSOM ge馱perimenteerd. De methode werkt nu reeds voor 7.000.000 documenten. Verder is WEBSOM, ge馱lementeerd op een parallelle computer waarbij de units (neuronen) van het tweedimensionale netwerk worden verdeeld over de processoren, goed *schaalbaar*. In een iteratief (leer-)proces worden de modelvectoren voor de woordfrequentie-vectoren in de units door middel van plaatsafhankelijke weging zodanig aangepast, dat een ruimtelijk patroon ontstaat dat een afspiegeling vormt van de hoog-dimensionale puntenwolk. Aan het

ruimtelijke patroon kunnen automatisch labels worden toegekend.

Aan de Technische Universiteit van Wenen wordt soortgelijk onderzoek gedaan, met als toepassing ondermeer de classificatie van elektronische bibliotheken.

Tot slot

Behalve voor de visualisatie van documenten, wordt het Kohonen-algoritme ook toegepast voor de visualisatie, de classificatie, het vinden van uitzonderingen etc. van vele andere hoog-dimensionale dataverzamelingen. Denk bijvoorbeeld aan *data mining*.

Omdat de natuur zich zijn geheimen, zoals in dit geval de werking van de hersenen, niet zomaar laat ontfutselen, zijn de toepassingen van het Kohonen-algoritme niet altijd zonder problemen uit te voeren. Het aantal berekeningen en iteraties is voor 'echte' problemen erg hoog, waardoor een *high performance* computer dikwijls gewenst is; een dergelijke computer is evenwel op het RC aanwezig.

Verder liggen verwante documenten of andere hoog-dimensionale data soms niet aaneengesloten maar in verspreide gebieden. Dit neemt echter niet weg, dat WEBSOM en soortgelijke visualisatiemethoden naar mijn overtuiging essenti馱 hulpmiddelen zullen worden bij het speuren op het web naast de woordgeori馱nteerde zoekmachines.

Raadpleeg de webversie van dit artikel www.rug.nl/rc/organisatie/pictogram/archief/hpc/projects/docvis voor links naar literatuur en vrije software.

▲ Begin pagina ▲

[index](#) Pictogram 3