Margot Gerritsen
Computational and Mathematical Engineering
Stanford University, CA, USA
gerritsen@stanford.edu

David F. Gleich
Informatics
Sandia National Labs, Livermore, CA, USA
dfgleic@sandia.gov

Ying Wang
Computational and Mathematical Engineering
Stanford University, CA, USA
yw1984@stanford.edu

Xiangrui Meng
Computational and Mathematical Engineering
Stanford University, CA, USA
mengxr@stanford.edu

Farnaz Ronaghi
Management Science and Engineering
Stanford University, CA, USA
farnaz@stanford.edu

Amin Saberi
Management Science and Engineering
Stanford University, CA, USA
saberi@stanford.edu

Onderzoek

Licht in de digitale duisternis dankzij computertools voor digitaal beheer

Computertools zijn niet meer weg te denken bij tegenwoordige zoek- en aanbevelingstechnologieën. Moderne digitale archieven bestaan echter uit ongekend gevarieerde collecties van gedigitaliseerd materiaal en zogenaamde *born-digital content*. Het is nog altijd lastig om interessant materiaal in deze archieven op te zoeken. Vaak ontbreken hierin annotaties — of metagegevens — op basis waarvan mensen het interessantste materiaal kunnen vinden. David F. Gleich, Ying Wang, Xiangrui Meng, Farnaz Ronaghi, Margot Gerritsen en Amin Saberi van Computational Approaches to Digital Stewardship (CADS) werken aan een visie op een virtueel bibliotheeksysteem waarmee makkelijker de interessantste parels kunnen worden gevonden in gevarieerde collecties van digitale archieven. Zij beschrijven vier computertools die zij hebben ontwikkeld zodat digitale archieven beter kunnen worden verwerkt en onderhouden. De eerste tool is een verbeterd algoritme voor de indeling van grafen met honderdduizenden knooppunten. De tweede tool is een nieuw algoritme voor het afstemmen van databases met koppelingen tussen de objecten, ook bekend als netwerkalignmentprobleem. De derde tool is een heuristische optimalisatiemethode waarmee een reeks geografische verwijzingen in een boek worden gedesambigueerd. En de vierde tool is een techniek waarmee automatisch een titel wordt gegenereerd op basis van een beschrijving.

In de afgelopen 25 jaar is het karakter van documenten in onze samenleving veranderd. Voorheen werden documenten op papier of op een ander fysiek medium opgeslagen. Tegenwoordig worden onze documenten opgeslagen in digitale bestanden. Deze situatie stelt ons voor een subtiel probleem. Bedenk eens hoeveel van uw eigen — digitaal opgeslagen — werk niet langer toegankelijk is omdat:

– het programma waarin het bestand moet worden gelezen, niet meer beschikbaar is;
– het programma waarin het bestand moet worden gelezen, niet meer werkt met oude bestanden;
– er geen hardware meer bestaat om de fysieke media te lezen.

Kuny [30] zet de basis voor het probleem uiteen en bedacht de uitdrukking *een digitale duisternis* om de ernst van de situatie duidelijk te maken. Ook beschrijft hij enkele oplossingen die nodig zijn om dit aan te pakken. Deze ideeën zijn grotendeels gericht op het probleem om digitale bits, opslag en bestandsindelingen te behouden. Zo heeft Kuny als interessante uitdaging vastgesteld dat digitale opslag een openbaar goed moet worden. We zijn afhankelijk van historische documenten uit het verleden om het heden te informeren. Daarom moeten onze documenten voor dit doel behouden blijven. Het probleem met het bewaren van documenten is dat dit alleen nut heeft wanneer de informatie door iemand wordt gebruikt. Voor de meest succesvolle opslagactiviteiten moeten de gegevens dus beschikbaar en eenvoudig toegankelijk worden gemaakt.

Uitdagingen in digitale webarchieven

Alleen al het bieden van toegang tot de gegevens zorgt voor de nodige uitdagingen. Van oudsher was materiaal opgeslagen in een bibliotheek en gingen wetenschappers *naar* de bibliotheek om dit in te kijken. Eenmaal daar overlegden ze met archivarissen om te bepalen welk materiaal ze precies nodig hadden. Tegenwoordig verwachten gebruikers toegang vanaf elk apparaat met een internetverbinding. Eigenlijk — en misschien vooral als reactie op de efficiënte zoekmachine van Google — verwachten we een direct antwoord op onze slecht geformuleerde informatieverzoeken. Het probleem met een dergelijke werkwijze in deze digitale collecties is dat gebruikers vaak iets willen ontdekken in plaats van opzoeken. Met andere woorden, ze willen niet met systemen iets zoeken wat ze al weten, maar iets nieuws vinden wat ze interessant vinden. Zo zou het volgende gesprek in een bibliotheek kunnen hebben plaatsgevonden:

Bibliothecaris Kan ik u ergens mee helpen?
Bezoeker Ik ben onlangs vanuit Zweden hierheen verhuisd. Hebt u ook een goed boek over lokale geschiedenis?
Bibliothecaris Oh, veel van onze eerste immigranten kwamen uit Zweden. Ik heb precies het juiste boek voor u.

Onze hoop is dat we dergelijke hulp in een digitaal archief kunnen bieden. Laten we eens bekijken hoe dit scenario online zou kunnen verlopen om te begrijpen welke uitdagingen zich voordoen bij het bieden van toegang tot digitale archieven.

Gebruiker Voer een zoekopdracht voor 'lokale geschiedenis' in.
Systeem Geef een ranglijst met antwoorden weer om aan te geven wat de beste naslagwerken zijn voor informatie over lokale geschiedenis; samen met een lijst met belangrijke subonderwerpen, zoals Zweedse immigranten.
Gebruiker Klik op de lijst met subonderwerpen over Zweedse immigranten.
Systeem Geef een nieuwe ranglijst met antwoorden weer, waarvan er één is gemarkeerd als 'speciaal belichte selectie'.

Bedenk welke technologieën nodig zijn voor deze interactie. Ten eerste moet een dergelijk systeem weten dat de zoekopdracht 'lokale geschiedenis' verwijst naar de geschiedenis van de regio waar de zoeker zich bevindt, of dat deze een specifieke lokale geschiedenis impliceert. Ten tweede moet de zoekmachine in staat zijn om te zoeken naar het onderwerp of trefwoorden die verband houden met elk item in de collectie. Ten derde is een procedure nodig om de resultaten te classificeren zodat een nuttige ranglijst kan worden teruggestuurd naar de gebruiker. Ten vierde moet binnen de zoekopdracht een reeks subonderwerpen worden vastgesteld.

Voor boeken verloopt dit vrij goed via bestaande tools. Ook hebben veel bibliotheken hun *openbare webcatalogi* oftewel OPAC's (Online Public Access Catalogs) herzien om dergelijke zoekopdrachten mogelijk te maken. Raadpleeg de websites van de bibliotheek van de North Carolina State University, Queens en Stanford, bijvoorbeeld:

http://www.lib.ncsu.edu/summon
http://www.queenslibrary.org
http://searchworks.stanford.edu

Figuur 1 Een foto van de immigratiekaart van John van Neumann (Johann von Neumann), gemaakt in de Library of Congres in januari 2007. Wetenschapshistorici zouden dit artefact graag willen *ontdekken*, maar weten niet hoe ze hiernaar moeten zoeken.

De informatie over onderwerpen in een boek worden vaak verstrekt via de LCSH-descriptors (Library of Congress Subject Heading). Voor boeken die in de Verenigde Staten zijn gepubliceerd, zijn de LCSH-descriptors te vinden op de eerste paar pagina's van veel boeken met de catalogusgegevens van de Library of Congress. Zo heeft het boek *Handbook of Writing for the Mathematical Sciences* [18] van Nick Higham de volgende onderwerptitels: 'Mathematics–Authorship' en 'Technical writing' Hieruit blijkt dat het boek gaat over de problemen met het schrijven van wiskundige formules en met technisch schrijven. Deze descriptors waren een oud soort indexering die werd toegepast op boeken zodat onderwerpen konden worden opgezocht in kaartencatalogi. De ruimte van een kaartencatalogus was beperkt. Daarom moest het mogelijk zijn om met zo min mogelijk indexingangen allerlei onderwerpen op te nemen in de indexering. Recenter is het ook mogelijk om via *full text*-zoekopdrachten boeken op te zoeken omdat er steeds meer *born-digital*-inhoud beschikbaar is en boeken op grote schaal worden ingescand. Gezamenlijk ondersteunen deze technologieën dergelijke zoekopdrachten voor boeken, maar er is nog ruimte voor toekomstige verbeteringen. Zo is de bovenstaande zoekopdracht voor 'lokale geschiedenis' vooral problematisch omdat 'lokale geschiedenis' een specifiek soort geschiedenis is die wordt beschreven in de onderwerptitels van de Library of Congress. Met een dergelijke zoekopdracht op deze systemen worden meestal boeken over het concept 'lokale geschiedenis' opgehaald. Een zoekresultaat was een boek over hoe u meer informatie over de geschiedenis van uw regio te weten kunt komen, dus geen boeken over de geschiedenis van de regio zelf.

Digitale opslag gaat echter veel verder dan boeken of gedigitaliseerde boeken. Het omvat zowel monumentale als alledaagse digitale artefacten. Voor dergelijke objecten zijn waarschijnlijk geen gegevens over onderwerptitels beschikbaar. Bovendien bestaan de items zelf mogelijk niet uit tekst. De Library of Congress heeft meer dan 14 miljoen afbeeldingen (volgens de webpagina van de bibliotheek: http://www.loc.gov/rr/print, geraadpleegd op 13 augustus 2010). Andere mogelijkheden zijn: enquêteresultaten, kaarten, audio en video. In de volgende hoofdstukken neemt het ontbreken van tekstbeschrijving van deze soorten materiaal een belangrijke plaats in, want het is niet altijd duidelijk hoe we gebruikers het beste in staat kunnen stellen om interessante artefacten te ontdekken. Onze huidige technieken zijn erop gericht om gegevens te extraheren uit de weinige tekst die we mogelijk over het item hebben.

Digitale archieven voor historisch materiaal

Tot nu toe hebben we het probleem rond de toegang tot digitale archieven beredeneerd vanuit het oogpunt van digitale opslag. In bibliotheken worden echter ook vele zeldzame, cultureel belangrijke manuscripten, foto's en andere objecten bewaard. Deze items zijn vaak kwetsbaar en niet geschikt om door

allerlei handen te gaan; en toch is de missie van een bibliotheek om deze items te delen. Via digitalisering en beeldverwerking wordt een doeltreffende kopie verkregen die op brede schaal kan worden gedeeld. Dezelfde moeilijkheden doen zich echter voor wanneer mensen toegang krijgen tot deze items, zoals bij algemene digitale archieven. Laten we een voorbeeld geven. Tijdens een bezoek aan de manuscriptafdeling van de Library of Congress wees een van de inhoudsdeskundigen ons op een doos met artefacten van John von Neumann. Een van deze items was een kopie van zijn immigratiekaart (zie Figuur 1). Digitale opslag is bedoeld om interessant materiaal in een breed en gevarieerd archief te kunnen vinden. Op dezelfde manier zijn deze speciale digitale collecties bedoeld om parels te kunnen vinden, zoals — wat ons betreft — informatie over John von Neumann. We zouden niet weten hoe we anders zelf hiernaar hadden moeten zoeken.

Dit is inmiddels een acuut probleem in de Library of Congress. Rond 1994 startte deze bibliotheek een enorm project om enkele van de belangrijkste werken uit de Amerikaanse cultuur te digitaliseren. Het resultaat was de collectie *American Memory* met een webinterface. Tot de gedigitaliseerde collecties behoren het dagboek van George Washington, brieven van Abraham Lincoln, en de eerste films die door Thomas Edison zijn opgenomen. Het was echter moeilijk om mensen bij het materiaal in deze collectie te krijgen. Hoewel tijdens de eerste digitalisering enkele beperkte metagegevens werden verzameld, waren deze activiteiten vooral gericht op digitalisering in plaats van effectieve toegang tot het materiaal. Bijna twintig jaar later wilde de Library of Congress deze collecties aanpassen aan moderne standaarden voor digitale archieven. Hiermee bedoelen we toegangspatronen zoals hierboven. Hiervoor zijn in elk geval accurate metagegevens over onderwerp, plaats, tijd en mensen nodig.

Historisch gezien werden deze metagegevens door bibliothecarissen of inhoudsdeskundigen gemaakt. Omdat digitalisering tegenwoordig echter zo eenvoudig is, kunnen de deskundigen de hoeveelheid materiaal niet bijhouden om dit te annoteren. De UNESCO heeft onlangs de Digitale wereldbibliotheek opgezet om te proberen de belangrijkste artefacten ter wereld op te nemen in een digitaal webarchief. De grootte van de oorspronkelijke collectie werd beperkt omdat de UNESCO behoefte had aan goed georganiseerde metagegevens die handmatig werden vertaald in elk van de zeven talen van de VN. Moet onze toegang tot deze artefacten worden beperkt doordat deskundigen voor de lastige taak staan om alles te annoteren en vertalen?

OPAC	Online Public Access Catalog
MARC	MAchine Readable Cataloging
XML	eXtensible Markup Language
RDF	Resource Description Framework
LCSH	Library of Congress Subject Headings
HIT	Human Intelligence Task
born-digital	inhoud die altijd alleen maar in digitale vorm heeft bestaan
artefact	object in een digitaal archief
metagegevens	informatie *over* een digitaal object, met name *tijd*, *plaats* en *onderwerp*
crowd-sourced	een term waarmee gegevens worden beschreven die zijn verzameld uit officieuze bronnen
folksonomie	een specifiek type crowd-sourced gegevens met een reeks tags — korte beschrijvingen — die zijn toegepast op een reeks objecten in een database
tags	laagste niveau van een folksonomie

Tabel 1 Afkortingen en definities.

Overzicht

Laten we kort aangeven welke problemen zich voordoen bij het opbouwen van zoek- en bladertools in deze archieven. Ten eerste zijn de items zeer heterogeen: boeken zijn slechts een klein gedeelte van de collectie die kan worden doorzocht. Ten tweede zijn de metagegevens voor alles (behalve voor boeken) inconsistent en onvolledig, terwijl de nuttigste metagegevens mogelijk niet beschikbaar zijn. Ten derde bestaan er geen systeemeigen koppelingen tussen items. Ten vierde is de inhoud opgesteld in veel talen. Ten vijfde is het lastig om deze items te classificeren vanwege de zeer inconsistente metagegevens.

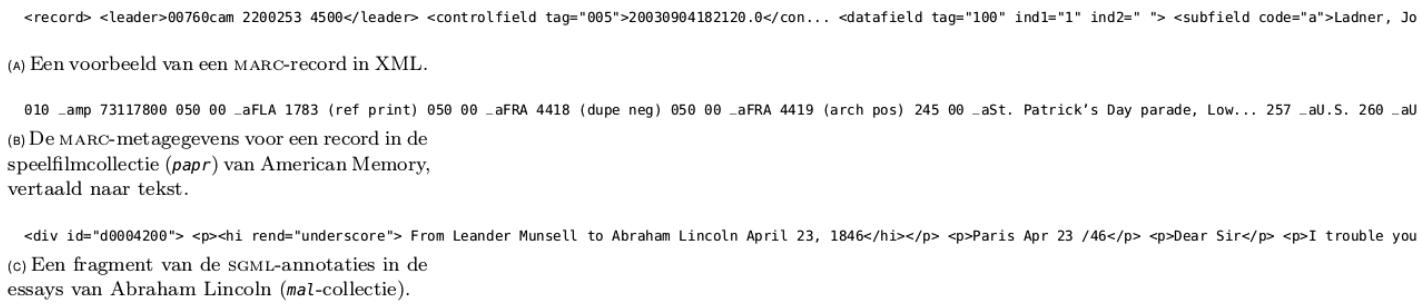
In dit artikel geven we geen uitgebreide oplossing voor deze problemen. In plaats hiervan halen we kleine, handelbare en interessante computerproblemen uit de visie op onze digitale bibliothecaris.

Hieronder beschrijven we enkele problemen uit ons onderzoek.

Als eerste probleem bespreken we de mengelmoes van beschikbare gegevens. Zie Tabel 1 voor een overzicht van de gegevens die we willen opzoeken en de gegevens die we kunnen gebruiken bij de zoekopdracht. We beschrijven elke gegevensset uitgebreider in de volgende paragraaf. Hoewel we als allesomvattend doel een uniforme zoek- en bladerinterface mogelijk willen maken, zijn de objecten waarnaar we willen zoeken en bladeren, divers. Een ander probleem is dat sommige gegevens die we mogelijk willen gebruiken, behoorlijk gecompliceerd zijn. Zo zijn de onderwerptitels van de Library of Congress een thesaurus waarmee een onderwerp uniek wordt geïdentificeerd. Deze wordt al meer dan honderd jaar gebruikt. Er worden volledige cursussen over deze database gegeven in curricula van informatiewetenschappen. Hoe kunnen we dan snel meer hierover te weten komen? Ons antwoord is visualisatie en we gaan in in de volgende paragraaf dieper op deze werkwijze in.

Toen we de structuur van de onderwerptitels van de Library of Congress eenmaal begrepen, viel het ons op dat deze verwant was aan de structuur van de categorieën die aan Wikipedia ten grondslag liggen. Naar aanleiding hiervan hebben we onderzocht hoe we de onderwerptitels van de Library of Congress konden *afstemmen* op de categorieën in Wikipedia. En bovendien hebben we hierdoor nagedacht over andere bronnen van *openbare* of *crowd-sourced* gegevens. In de paragraaf 'Openbare crowd-sourced gegevens' bespreken we ons idee om de onderwerptitels van de Library of Congress af te stemmen op Wikipedia-categorieën. Ook gaan we in op uitdagingen bij het gebruik van deze gegevenstypen.

Op dit punt doet zich een belangrijk probleem voor. Zoals we hebben opgemerkt, willen we vaak gegevens over de *plaats* en het *onderwerp* van elk object in onze collectie. Deze gegevens zijn echter niet altijd beschikbaar. In de volgende twee paragrafen stellen we ideeën voor om deze *ontbrekende metagegevens* te genereren. In de paragraaf 'Ambigue geografische verwijzingen' introduceren we een optimalisatieprobleem om geografische plaatsnamen te desambigueren. Er wordt dus geprobeerd de volgende vraag te beantwoorden: verwijst 'San Jose' naar San



```
<record> <leader>00760cam 2200253 4500</leader> <controlfield tag="005">20030904182120.0</con... <datafield tag="100" ind1="1" ind2=" "> <subfield code="a">Ladner, Jo
```

(A) Een voorbeeld van een MARC-record in XML.

```
010 _amp 73117800 050 00 _aFLA 1783 (ref print) 050 00 _aFRA 4418 (dupe neg) 050 00 _aFRA 4419 (arch pos) 245 00 _aSt. Patrick's Day parade, Low... 257 _aU.S. 260 _aU
```

(B) De MARC-metagegevens voor een record in de speelfilmcollectie (*papr*) van American Memory, vertaald naar tekst.

```
<div id="d0004200"> <p><hi rend="underscore"> From Leander Munsell to Abraham Lincoln April 23, 1846</hi></p> <p>Paris Apr 23 /46</p> <p>Dear Sir</p> <p>I trouble you
```

(C) Een fragment van de SGML-annotaties in de essays van Abraham Lincoln (*mal*-collectie).

Figuur 2 Drie voorbeelden van onze gegevensbestanden. Deze blik op het binnenste van elk bestand toont hoe de bibliotheek records er in ongemaakte vorm uitzien.

Jose in Californië of naar San Jose in Costa Rica? In de paragraaf 'Metagegevens en titelremediatie' beschrijven we hoe automatisch een titel en een reeks trefwoorden kunnen worden gegenereerd op basis van een tekstbeschrijving.

We sluiten af met een samenvatting en richtingen voor toekomstig onderzoek.

Gegevens begrijpen

Zoals we hebben opgemerkt in de inleiding, is onze visie op een virtuele bibliothecaris vrijwel grenzeloos. Een van de gevolgen van deze visie is dat we allerlei bestaande gegevensbronnen moeten verwerken. Al deze gegevenssets hebben weer een andere indeling. Soms hebben ze zelfs niets met elkaar te maken. Desalniettemin is ons doel om de gegevens samen te voegen en onze virtuele bibliothecaris mogelijk te maken door bijvoorbeeld de onvolledige metagegevens van een bibliotheekrecord aan te vullen met gegevens uit openbare bronnen. Tabel 2 bevat een overzicht van de verschillende gegevenssets die we in dit document gebruiken. Er zijn drie algemene groepen:

1. eigen gegevens van Library of Congress,
2. openbare en crowd-sourced gegevens,
3. meertalige gegevens.

De eerste groep bevat informatie die de Library of Congress meestal niet deelt, zoals de onopgemaakte metagegevens achter de collectie American Memory, of informatie die de bibliotheek verkoopt om de kosten terug te verdienen. De tweede categorie bestaat uit gegevens die volledig vrij beschikbaar zijn. We vertellen meer over deze categorie in de paragraaf 'Openbare crowd-sourced gegevens'. De laatste categorie is ook eigendom van de Library of Congress, maar onderscheidt zich doordat de metagegevens beschikbaar zijn in meerdere talen. In dit document richten we ons op de eerste twee categorieën, maar we bespreken ideeën voor de meertalige gegevens in de paragraaf over toekomstig werk. We willen nadrukkelijk erop wijzen dat deze lijst met gegevensbronnen niet volledig is. Er zijn veel andere bronnen die we hadden kunnen gebruiken. Deze lijst bevat alleen maar bronnen die *wij* hebben gebruikt.

In elk van deze databases of collecties wordt informatie op een eigen manier opgeslagen, waarbij zelfs binnen een collectie verschillen bestaan. American Memory is in feite een collectie van collecties. Sommige metagegevens in verband met de items hebben de MARC-indeling; andere zijn in de XML-indeling. Figuur 2 bevat een voorbeeld van enkele onopgemaakte gegevens in deze databases. De details van de MARC- [44], RDF- en XML-indeling zijn niet relevant. Elke gegevensindeling biedt globaal een reeks records en velden over deze records. Tenslotte kunnen sommige items annotaties in nog een andere indeling bevatten. Bij de *mal*-collectie zijn bijvoorbeeld metagegevens in XML-bestanden en annotaties in SGML-bestanden (een voorganger van XML) opgeslagen. We noemen al deze details en gege-

Type	Collectie	Aantal objecten	Indeling	Opmerkingen
Eigendom van Library of Congress	Onderwerptitels	298.964	MARC	Autoriteitsbestanden van dec. 2006
	Naamautoriteiten	6.662.688	MARC	Autoriteitsbestanden van dec. 2006
	Catalogus	7.207.747	MARC	Boekencatalogus van Library of Congress
	American Memory	617.673	MARC of XML	101 heterogene collecties
	— *papr*	703	MARC	Speelfilms
	— *mal*	20.158	XML	Essays van Abraham Lincoln
	— *gmd*	6888	MARC	Kaartencollectie
	— *wpa*	2000	XML	American Life Histories
Openbaar en crowd-sourced	Wikipedia	3.799.337	XML	(Vanaf april 2007)
	Wikipedia-categorieën	226.221	(afgeleid)	(Vanaf april 2007)
	Geografische namen	6.914.549	Tekst	Een geografisch woordenboek
	Project-Gutenberg	24	Tekst	Tekstboeken
Meertalig	Global Gateways	21.274	MARC of tekst	
	Digitale wereldbibliotheek	196	XML	

Tabel 2 Een overzicht van de gebruikte gegevens tijdens ons onderzoek. Voor elke collectie vermelden we de grootte als het aantal 'dingen' in de collectie. American Memory is een groep collecties. *papr*, *mal*, *gmd* en *wpa* zijn dus subcollecties binnen American Memory.

vensindelingen om te benadrukken hoe heterogeen de onopgemaakte gegevens zijn, zelfs op het laagste niveau. We moeten doorlopend nieuwe interpreters voor elk van deze gegevenscollecties schrijven om eenvoudigweg de gegevens zelf te kunnen openen.

Nadat we de gegevens hebben geopend, stapelen de problemen zich op. In een ideale wereld zou elk item een volledige reeks consistent gespecificeerde metagegevens bevatten, inclusief datum, locatie, onderwerp en personen. De werkelijkheid laat echter veel te wensen over. In de paragraaf 'Metagegevens en titelremediatie' zullen we zien hoe inconsistent sommige metagegevens binnen deze bestanden zijn. Zodra we de gegevensbestanden kunnen lezen, doet zich echter een ander probleem voor: we moeten de inhoud begrijpen. Met begrijpen bedoelen we dat we vertrouwd moeten zijn met de bijzondere kenmerken van een gegevensset: idealiter zoals een deskundige die al jarenlang met de gegevens werkt. Zoals we eerder hebben opgemerkt, zijn sommige van deze gegevens in de afgelopen honderd jaar verzameld door de bibliotheek [4]. In de volgende subparagraaf duiken we in de onderwerptitels van de Library of Congress om te laten zien hoe we inzicht kunnen krijgen in de inhoud van deze databases.

Een grafische weergave van LCSH

De Library of Congress Subject Headings (LCSH) is een database van termen die worden bijgehouden door de Library of Congress en worden gebruikt voor de indexering van het onderwerp van bibliografische documenten en ook voor kruisverwijzingen tussen gerelateerde onderwerpen. Een onderwerptitel bevat bredere termen, smallere termen en 'zie ook'-termen.

De onderwerptitel 'Mathematics' houdt bijvoorbeeld verband met de bredere term 'Science' en de smallere termen 'Algebra', 'Economics', 'Mathematical' en 'Women in mathematics'. We kunnen de LCSH-database zien als een ongerichte graaf waarin elke onderwerptitel een knoop is en met elke relatie een ongerichte lijn wordt gedefinieerd.

Omdat we snel inzicht wilden krijgen in de informatie binnen deze verbindingen, wilden we de graaf visualiseren. Een grote graaf kan vaak op twee manieren worden gevisualiseerd: (i) in kleine delen [37]; of (ii) als volledige graaf. We hebben met beide technieken gewerkt en beschrijven alleen de tweede om ruimte te besparen. Door de volledige graaf te visualiseren krijgen we inzicht in de algehele verbindingsstructuur van het netwerk zodat we mogelijk gerichtere vragen kunnen stellen. Zie [21] voor inzichten in het Twitter-netwerk, als voorbeeld van dit type analyse. Als we een graaf met honderdduizenden knopen willen visualiseren, hebben we een middel nodig om een indeling (een toewijzing van punten aan coördinaten in het vlak) te berekenen. Dit is een uitdagende berekening waarnaar nog altijd onderzoek wordt verricht (zie [23] voor een recente bijdrage over de visualisatie van grote grafen).

We hebben het LGL-programma (Large Graph Layout) [1] gebruikt, dat ongeveer als volgt werkt:

1. zoek een minimale spanning-tree voor de graaf;
2. zoek de knoop met de minimale totale afstand van kortste paden naar alle andere knopen (we noemen deze knoop het midden);
3. tel voor $k = 1$ tot ... alle knopen op die op een afstand van k lijnen vanaf de middelste knoop liggen, en optimaliseer lokaal hun positie.

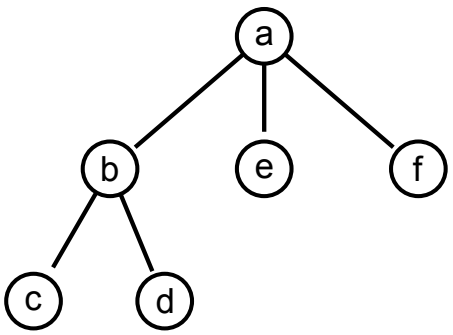
We kiezen LGL vanwege stap 1. Tijdens voorbereidend werk met de graaf achter LCSH ontdekten we dat grote delen hiervan een boomstructuur kunnen hebben. Daarom zou een indelingsalgoritme waarmee de boomstructuur in de graaf wordt onderzocht, een nuttige structuur moeten opleveren.

Tijdens het LGL-proces gaat het meeste werk in de laatste stap zitten: er moet een dynamische simulatie worden uitgevoerd om globaal een minimale energietoestand te berekenen. Zie het essay over LGL voor meer informatie over stap 3; wij richten ons op stap 2. Toen we met de code aan de slag gingen, duurde het ongeveer twee uur om het midden te vinden. In de oorspronkelijke implementatie van het LGL-algoritme werd de boomstructuur niet gebruikt bij het berekenen van de middelste knoop. Zoals we zo zullen zien, is het totaal van alle kortste paden voldoende voor een eenvoudige herhaling in een boomstructuur. We implementeerden een procedure om het midden efficiënt te berekenen. Nadat we deze wijziging hadden doorgevoerd, duurde het maar enkele seconden om stap 2 te berekenen.

We beschrijven nu onze optimalisatie om het midden efficiënt te berekenen. Stel dat $D_{u,v}$ het aantal lijnen langs het kortste pad tussen knoop u en v in de minimale spanning-tree is. We zijn op zoek naar de knoop c waarvoor $\sum_v D_{c,v}$ zo laag mogelijk is. Het belangrijkste idee achter onze optimalisatie is dat er altijd een uniek pad tussen twee knopen in een boomstructuur loopt. Formeel is de procedure als volgt. Neem $C_u = \sum_v D_{u,v}$ als de score voor het midden van knoop u. Wijzig in een boomstructuur eventueel C_u in C_w wanneer we een lijn hebben (u, w). We hoeven alleen maar te berekenen hoeveel paden die beginnen bij u, langer worden wanneer ze in plaats hiervan beginnen bij w, en hoeveel paden korter worden wanneer ze beginnen bij w. Zie Figuur 3 voor een voorbeeld van de manier waarop we van deze waarneming kunnen profiteren. Kies een willekeurige knoop a en laat de boomstructuur beginnen bij a ('root') om deze waarneming te implementeren voor een structuur T. Bereken vervolgens C_a. Voor elke knoop w die is verbonden met knoop a, geldt:

$$C_w = C_a - N_w + (n - N_w).$$

Hierin is N_w het aantal knopen in de substructuur dat begint bij w, en is n het totaal aantal knopen. Misschien wordt de formule duidelijker als u bedenkt dat alle paden vanaf w naar knopen in de subboomstructuur die beginnen bij w, één lijn korter zijn. Daarom verlagen we C_a ook met N_w. Bovendien bevatten alle overige knopen in de graaf ($(n - N_w)$ in totaal) paden die één lijn langer zijn wanneer ze beginnen bij w in plaats van a. Door deze procedure voor alle verdere niveaus te herhalen, kunnen we C_v voor elke knoop v berekenen in lineaire tijd. Voor het volledige proces moet de graaf driemaal worden doorlopen: eerst om C_a te berekenen voor de arbitraire root; dan om de grootte van elke subboomstructuur te berekenen (N_v voor elke v); en tenslotte om C_v te berekenen, waarbij C_a en N_v voor elke knoop zijn gegeven.



Figuur 3 Een klein voorbeeld van de manier waarop we de totale afstand van de kortste paden naar alle knopen efficiënt berekenen in een boomstructuur. We kunnen eenvoudig de totale afstand van de kortste paden berekenen wanneer we beginnen bij de valse root a: C_a=7. Als we nu C_b willen berekenen, zien we dat er drie knopen een lijn verder weg komen te liggen (a,e,f) en drie knopen één lijn dichterbij komen te liggen (b,c,d). Dus C_b=C_a−3+3=7. Evenzo vinden we dus C_c=C_b−1+5=11, en hetzelfde geldt voor C_d. Zowel C_e als C_f zijn ook 11. In dit voorbeeld kan a of b de rootknoop zijn. Ook N_a=7, N_b=3 en N_c=N_d=N_e=N_f=1.

Nadat we deze wijzigingen hadden aangebracht, voerden we het LGL-algoritme uit op de grootste verbonden component van de ongerichte graaf van LCSH. Figuur 8 aan het eind van dit artikel bevat een visualisatie waarin de door LGL berekende indeling wordt gebruikt. Lijnen zijn getekend met alfamenging om de lokale dichtheid te laten zien. Elk knooppunt is gekleurd op basis van een clustering die via het CLUTO-programma [24] is berekend. We zien grote vlakken met dezelfde kleur. Dit betekent dat met zowel CLUTO als LGL soortgelijke structuren in de graaf worden geïdentificeerd. Zie voor een uitgebreidere kijk op deze visualisatie

http://cads.stanford.edu/lcsh-galaxy

Op basis van deze visualisatie vinden we de volgende structuur in het LCSH-netwerk. Er is een dichte kern van onderwerptitels van algemeen belang, zoals 'Law', 'Science' en 'Art'. Rond deze kern zien we een aantal meer esoterische onderwerpen, inclusief een uitgebreide regio met geografische onderwerpen ten zuiden van de gele kern. Een ander inzicht is dat mogelijk niet alle regio's even goed zijn gecategoriseerd. Links op de afbeelding bevindt zich een grote, stervormige constructie rond de onderwerptitel Japan-Antiquities. Deze ster bevat meer dan duizend onderwerptitels, met slechts één verbinding terug naar het midden van de ster. Andere regio's van de graaf (zoals de onderwerptitels voor talen linksboven) blijken daarentegen beter te zijn georganiseerd.

Nadat we veel te lang deze visualisatie hadden bestudeerd en hiernaar hadden zitten staren, hadden we het gevoel dat we meer inzicht hadden in de onderwerptitels van de Library of Congress. In feite deden enkele kenmerken van de graaf ons denken aan een andere graaf: de categoriestructuur van Wikipedia. In het volgende hoofdstuk gaan we in op deze relatie nadat we kort crowd-sourced gegevens hebben uitgelegd.

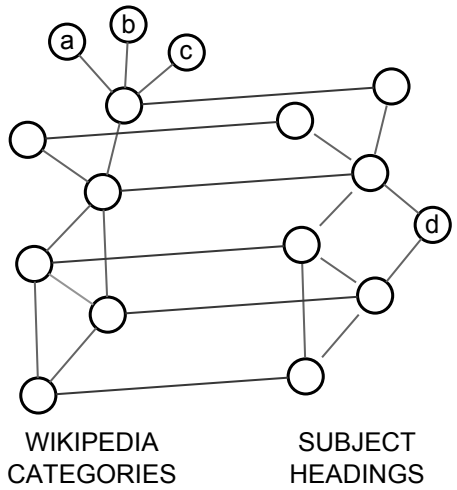
Openbare crowd-sourced gegeven

Ongeveer vijftien jaar geleden kostte een hoogwaardige encyclopedie een paar honderd euro. Tegenwoordig heeft iedereen met een internetverbinding gratis toegang tot Wikipedia. In feite mag de volledige inhoud van Wikipedia bulksgewijs worden gedownload voor niet-commercieel gebruik. De encyclopedie Wikipedia is misschien wel het beste voorbeeld van de tweede gegevenscategorie die we tijdens ons werk onderzoeken: openbare en crowd-sourced gegevens. Een openbare gegevensset is eenvoudig gezegd een gegevensset die gratis op internet beschikbaar is. Een voorbeeld van een openbare gegevensset is de website http://id.loc.gov/authorities, waar de onderwerptitels van de Library of Congress op een interactieve manier kunnen worden onderzocht en bulksgewijs kunnen worden gedownload. De records achter LCSH worden echter nog altijd beheerd door de Library of Congress.

Wikipedia is daarentegen een voorbeeld van crowd-sourced gegevens. In de afgelopen tien jaar is de encyclopedie zonder toezicht geschreven en bewerkt door uiteenlopende personen. Ze ontwikkelden een zelfregulerend mechanisme waarmee vrijwel iedereen een bijdrage kon leveren aan de encyclopedie, terwijl personen minder mogelijkheden hadden om de inhoud voor eigen doeleinden te bewerken. Let eens op het verschil met oude modellen voor informatieverzameling. Op gegevensopslagplaatsen werd toezicht gehouden door een 'gezegende' groep deskundigen die wijzigingen beoordeelden en autoriseerden in een poging om fouten te voorkomen. In het geval van LCSH duurde het proces tientallen jaren, waarbij de regels voor het toevoegen van nieuwe items alleen bekend waren binnen een select gezelschap. Zoals we straks zullen zien, zette Wikipedia een soortgelijk categoriesysteem in slechts enkele jaren op.

Crowd-sourcing is een ongekend succes. Het is een pijler van zogenaamde Web 2.0-technologieën geworden en speelde een prominente rol op de websites van Flickr en Delicious. Volgens een theorie die wordt aangehangen door Surowiecki [41], kunnen veel betrouwbaardere voorspellingen worden gedaan op basis van de uiteenlopende perspectieven van veel mensen dan van enkele deskundigen. Deze theorie staat bekend als de 'wijsheid van de menigte'. Uit een recent onderzoek naar folksonomieën, een veelgebruikt type crowd-sourced gegevens waarmee items met enkele korte tags worden beschreven, zoals op Flickr en Delicious, blijkt dat de tags van 'breedsprakige beschrijvers' nuttiger zijn dan de tags van 'categoriseerders' [28]. Als we ervan uitgaan dat deskundigen eerder tot de laatste categorie behoren, kan dit worden gezien als een empirische validatie van de methodologie voor crowd-sourcing.

Ongeacht de theoretische ondersteuning is er tegenwoordig een enorme hoeveelheid gegevens beschikbaar uit deze wat meer onsystematische modellen voor informatieverzameling. We vroegen ons het volgende af: kan een bibliotheek deze gegevens gebruiken

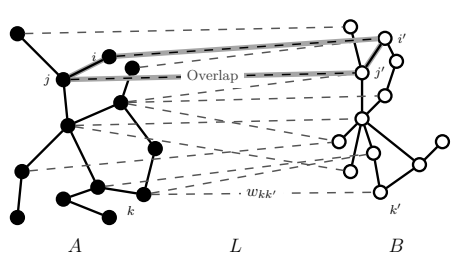


Figuur 4 We willen de Wikipediacategorieën afstemmen op de onderwerptitels van de Library of Congress. Via deze 1-op-1-afstemming (de horizontale verbindingen) kunnen we nieuwe onderwerptitels voorstellen (de knooppunten a, b, en c) en informatie toevoegen aan Wikipediacategorieën. (Het knooppunt d zou ons iets nuttigs moeten kunnen vertellen over de afgestemde buren.)

om onze virtuele bibliothecaris te implementeren? Het nut van deze openbare gegevensverzamelingen werd al vroeg erkend in [32]. We zullen in de rest van deze paragraaf laten zien hoe dit zou kunnen werken door de Wikipedia-categorieën af te stemmen op de onderwerptitels van de Library of Congress. In de volgende paragraaf bekijken we hoe openbare informatie rechtstreeks kan worden gebruikt om het probleem rond het zoeken naar de geografische verwijzingen in een document op te lossen.

Laten we nog eens teruggaan naar de structuur van de onderwerptitels van de Library of Congress: elk onderwerp is gerelateerd aan andere onderwerpen via de referenties 'Bredere term', 'Smallere term' en 'Zie ook'. We interpreteren deze relaties als een ongerichte graaf. De categoriepagina's in Wikipedia hebben een soortgelijke structuur.

Elke pagina in Wikipedia is lid van een of meer categorieën. Zo behoort de pagina over 'Singular Value Decomposition' tot de ca-



Figuur 5 Bij het algemene netwerkalignmentprobleem is het doel om de knopen van graaf A af te stemmen op de knopen van de graaf, terwijl B zoveel mogelijk lijnen probeert te *overlappen* en het gewicht van de lijnen in de overeenkomende $\sum w_{kk'}$ probeert te optimaliseren. Formeel overlapt een lijn bij een overeenkomst wanneer (i,j) een lijn in A is en het evenbeeld binnen de overeenkomst, $(m(i),m(j))$, ook een lijn in B is.

tegorieën 'Linear algebra', 'Matrix theory' en 'Functional analysis'. Categorieën hebben subcategorieën en gerelateerde categorieën die een hiërarchische structuur met enkele aanvullende lijnen vormen. Het lijkt misschien verrassend, maar de ongerichte graaf van Wikipedia-categorieën heeft ongeveer evenveel knopen als de LCSH-graaf: 205.948 tegenover 297.266. Andere kenmerken lijken ook op elkaar: de grootste verbonden component bestaat in beide grafen uit ongeveer 150.000 knopen, de gemiddelde afstand tussen knopenparen is ongeveer 7 in beide grafen, en ongeveer 6000 knooppunten hebben *identieke* tekstlabels.

Op basis van deze resultaten wilden we elke knoop in de LCSH-graaf *afstemmen* op of *toewijzen* aan een knoop in de Wikipediagraaf. We proberen een overeenkomst te vinden omdat de deskundigen die de LCSH hebben ontwikkeld, de overeenkomsten kunnen gebruiken om de dekking te verbeteren op nieuwe of snel ontwikkelende gebieden die mogelijk een betere dekking hebben in Wikipedia. Zie Figuur 4 voor een voorbeeld. We formaliseren het probleem als probleem bij het alignen van een licht ('sparse') netwerk [2]. Uit de oplossing blijkt hoe we de knopen van twee grafen moeten afstemmen wanneer we over een redelijke set *potentiële overeenkomsten* beschikken. Figuur 5 bevat de structuur van het netwerkalignmentprobleem. Voor deze kwestie stellen we ook in [2] een algoritme voor het doorgeven van berichten voor. Met ons algoritme worden binnen

LCSH	↔	Wikipedia
	goed	
Dollar, American (Coin)	↔	United States dollar coins
Web sites	↔	Websites
Environmentalists	↔	Environmentalists by nationality
Peninsulas–Southeast Asia	↔	Peninsulas of Asia
	vreemd	
Cosby family	↔	Bill Cosby Songs
Peasants in literature	↔	Peasant foods
	onzinnig	
Hot tubs	↔	Hot dogs
Masques	↔	Vampire: The Masquerade

Tabel 3 Resultaten van onze afstemming van LCSH op Wikipedia. Zie de discussie in de tekst.

subject	werkwoord	object
Singular Value Decomposition	*is in categorie*	Linear algebra
Singular Value Decomposition	*is in categorie*	Matrix theory
Linear algebra	*is gerelateerd aan*	Affine geometry
Linear algebra	*is een subcategorie van*	Algebra

Tabel 4

enkele minuten vrijwel optimale oplossingen voor het probleem bij het alignen van het LCSH- en Wikipedia-netwerk verkregen (zelfs indien geïmplementeerd in Matlab). Tabel 3 bevat enkele overeenkomsten die via deze werkwijze zijn vastgesteld. De overeenkomsten zijn ingedeeld in drie groepen: *goed*, *vreemd* en *onzinnig*. Alleen de goede overeenkomsten zijn juist. De vreemde reeks overeenkomsten klopt 'bijna' en verwijst naar gerelateerde, maar andere concepten. De onzinnige reeks is gewoon helemaal fout. Bij onze huidige formulering van het probleem worden geen 'strafpunten' toegekend voor het identificeren van een overeenkomst waar we niet zoveel aan hebben. Daardoor zitten onze resultaten vol valse overeenkomsten. In toekomstig werk hopen we een strafpuntensysteem op te nemen.

Hoewel we ons algoritme zo hebben ontworpen dat dit werkt met honderdduizenden knooppunten, zijn er andere succesvolle technieken om knopen in een graaf af te stemmen. Dit probleem doet zich voor bij patroonherkenning, zie [6] voor een overzicht van dat werk. Er zijn ook allerlei fraaie matrixproblemen die zich voordoen. Zie [3, 12–13, 35–36] voor voorbeelden.

Uit recente onderzoeken komt één feit naar voren: *eenvoudige* algoritmen presteren net zo goed als of beter dan gecompliceerde algoritmen in het geval van aanvullende gegevens. Zie [38] voor een voorbeeld van dit verschijnsel in het probleem met Netflix-aanbevelingen en [15] voor een gedegen kijk op de rol van gegevens bij computergebruik. De krachtige mogelijkheid om twee gegevensverzamelingen af te stemmen lijkt handig om meer gegevens op te nemen. Hoewel de afstemming van LCSH op Wikipedia de aanzet gaf tot deze discussie, zijn we van mening dat netwerkalignment een middel is om openbare en crowd-sourced gegevens te gebruiken.

In algemenere zin staat het probleem rond het combineren van gekoppelde gegevens bekend als ontologieafstemming of -alignment. Een ontologie is een set beweringen waarmee relaties in een gestructureerde vorm worden uitgedrukt. Ze worden vaak beschreven als een set beweringen met een subject, werkwoord en object. Laten we nog eens kijken naar de Wikipedia-categorieën die eerder in dit hoofdstuk zijn genoemd. In Tabel 4 is te zien hoe ze als ontologie zouden worden uitgedrukt.

Algoritmen voor ontologie-alignment omvatten vaak *divide & conquer*-methoden om een soortgelijke doelstelling te optimaliseren als bij onze werkwijze voor netwerkalignment [10, 19].

Ambigue geografische verwijzingen

In de vorige paragraaf hebben we een techniek gezien waarmee we twee gerelateerde gegevensverzamelingen konden samenvoegen. We bekijken nu opnieuw een specifiek gegevenstype dat onze virtuele bibliothecaris nodig heeft: geografische metagegevens. De geografische context van een item is een essentieel stuk metagegevens om interessante informatie te ontdekken. Geografische correlaties zijn weliswaar meestal toevallig, maar fascinerend. Bovendien bieden geografische gegevens een eenvoudige manier om door een verzameling te bladeren of twee verschillende artefacten in verband te brengen. Niet alle items in *American Memory* hebben echter betrouwbare metagegevens. Een van de problemen waarmee we werden geconfronteerd, was hoe we de geografische entiteiten moesten extraheren uit een boek of document.

Hiervoor gaan we ervan uit dat we de tekst beschikbaar hebben of op een bepaalde manier aan beschrijvende tekst kunnen komen (misschien via spraakherkenning, Optical Character Recognition of crowd-sourced tags). Met deze tekst moet als eerste stap een lijst met locatienamen worden geëxtraheerd uit de tekst. Een locatienaam is een specifiek type benoemde entiteit. Bij een tekstverzameling kan een NER (*Named Entity Recognizer*) worden aangepast zodat alleen de namen worden uitgevoerd van tekst die waarschijnlijk staat voor naam van een locatie. We gebruikten de gratis beschikbare Stanford NER [11]. We hebben nog een stuk informatie nodig: de feitelijke geografische coördinaten van een locatienaam. Een database met toewijzingen tussen geografische coördinaten en locatienamen wordt een geografische index genoemd. We gebruikten GeoNames als onze geografische index. GeoNames is een gra-

tis beschikbare verzameling van ongeveer 7 miljoen plaatsnamen en de lengte- en breedtegraad van elke locatie. Gezamenlijk hebben we een verzameling plaatsnamen uit de Stanford NER-software en een verzameling cöordinaten van GeoNames. We hebben bijna ons doel bereikt: het vinden van alle plaatsen die in een boek of document worden genoemd. Plaatsnamen zijn echter niet gekoppeld aan unieke locaties. Verwijst de term 'San Jose' naar de hoofdstad van Costa Rica of naar het hart van Silicon Valley? Het antwoord op deze vraag is het probleem rond geografische desambiguering. Voor het antwoord moeten we context gebruiken.

Laten we het probleem in een wiskundige formule gieten. Stel dat $X = (x_1, \ldots, x_n)$ een reeks genoemde locaties is, gerangschikt op hun positie in de tekst. Deze reeks is de uitvoer van de NER-software. Formeel ging de locatienaam x_i vooraf aan x_j als $i < j$. Voor elke locatie x_i veronderstellen we dat er een set $Y_i = (y_{i,1}, \ldots, y_{i,k})$ is met bestaande locaties die overeenkomen met de tekstverwijzing x_i. Deze sets Y_i horen bij alle overeenkomsten in de GeoNames-database voor een locatienaam x_i. We noemen de set met alle mogelijke kandidaten $\mathcal{C}$, en daarmee elke $y_{i,r} \in \mathcal{C}$. Bovendien veronderstellen we dat we een afstandsfunctie tussen elementen in $\mathcal{C}$ hebben. Zie D nu als de geodetische afstand tussen de lengte- en breedtegraad van elke locatie. Deze functie wordt $D : \mathcal{C} \mapsto \mathbb{R}$. Ons doel is om een *bestaande* verwijzing voor elke kandidaat te kiezen. Deze verwijzingen kunnen op een natuurlijke manier worden gekozen door de afstand tussen de genoemde locaties tot een minimum te beperken. Dit idee wordt omgezet in het optimalisatieprobleem:

$$\begin{aligned} &\text{minimize} && \sum_{i=1}^{n-1} D(z_i, z_{i+1}), \\ &\text{subject to} && z_i \in Y_i \quad \text{voor alle } i. \end{aligned}$$

Als we dit probleem willen oplossen, kunnen we een dynamisch programma gebruiken. Stel dat $f_{j,r}$ de optimale oplossing is van

$$\begin{aligned} &\text{minimize} && \sum_{i=1}^{j-1} D(z_i, z_{i+1}), \\ &\text{subject to} && z_i \in Y_i \quad \text{voor alle } i, \\ & && z_j = y_{j,r}. \end{aligned}$$

Dan

$$f_{j+1,r} = \min_{s \in Y_j} \left(f_{j,s} + D\left(y_{j,s}, y_{j+1,r}\right) \right).$$

Met $\min_{r \in Y_n} f_{n,r}$ wordt het oorspronkelijke probleem tot een minimum beperkt. Voor dit Greedy Algoritme werkt $d = \max_j |Y_j|$ voor elke berekening van $f_{j+1,r}$. Er zijn hoogstens d van dergelijke berekeningen voor elke j, dus het totale werk van het algoritme is aan de bovenkant begrensd met nd^2. In praktijk zou d redelijk klein moeten zijn, want de meeste geografische entiteiten zullen een vrijwel unieke identifier hebben.

Onze zorg met dit algoritme is dat dit eenvoudig op het verkeerde been kan worden gezet door een verwijzing met één afstand. Bekijk het volgende fragment:

> Een Britse vakantieganger werd naar San Juan in Puerto Rico in plaats van San Jose in Costa Rica gestuurd door haar reisbureau. Andere toeristen die naar San Jose, Costa Rica wilden, kwamen in San Jose, Californië terecht en moesten toen de weg naar San Jose vragen.

(Geraadpleegd op http://www.skyscanner.net/news/articles/2010/09/007959-destination-doppelgangers-same-name-different-country.html op 8 september 2010.) Met het bovenstaande algoritme wordt bevestigd dat de eindverwijzing naar 'San Jose' betrekking heeft op 'San Jose, Californië' omdat de afstand 0 is, hetgeen onjuist is. Dit kan eenvoudig worden opgelost door aanvullende paarsgewijze afstanden op te nemen. Neem het gegeneraliseerde probleem:

$$\begin{aligned} &\text{minimize} && \sum_{\substack{0<j-i\le T \\ 0\le i,j\le n}} D(z_i, z_j), \\ &\text{subject to} && z_i \in Y_i \quad \text{voor alle } i. \end{aligned}$$

Nogmaals, we kunnen dit probleem oplossen met een variatie op het vorige dynamische programma. We tonen de generalisatie voor $T = 2$ en merken op dat grotere afstanden eenvoudiger af te leiden zijn. Stel dat $f_{k,(r,s)}$ de optimale oplossing is van

$$\begin{aligned} &\text{minimize} && \sum_{\substack{0<j-i\le 2 \\ j\le k}} D(z_i, z_j), \\ &\text{subject to} && z_i \in Y_i \text{voor alle } i, \\ & && z_{j-1} = y_{k-1,r}, \\ & && z_j = y_{k,s}. \end{aligned}$$

Dan

$$\begin{aligned} f_{j+1,(r,s)} = \min_{w \in Y_{j-1}} \Big(& f_{k,(w,r)} \\ & + D\left(y_{k-1,w}, y_{j+1,s}\right) \\ & + D\left(y_{k,r}, y_{j+1,s}\right) \Big). \end{aligned}$$

Indeling	Voorbeeld
####-##	1601-15
####-####	1862-1863
[Month] #, ####	Dec. 1, 1793
btw. #### and ####	btw. 1755 and 1762
#### [Season]	1939 Spring
anno ####	anno 1668
##/##/##	03/02/64
###-?	184-?
Bunka # ie ####	Bunka 1 ie 1804
Guangxu ## ####	Guangxu 30 1904
#########	185000930
United States	United States

Tabel 5 Deze tabel bevat een indeling van een type datumpatroon en een voorbeeld van dat patroon. De patronen worden weergegeven in vier groepen: duidelijk, ambigu, andere kalenders en verkeerd. Bunka 1 verwijst naar het eerste jaar van het Bunka-tijdperk in Japan, dus het jaar 1804. Evenzo is Guangxu 30 het dertigste jaar van het Guangxu-tijdperk in China, dus 1904. We hebben 'between' afgekort als 'btw'. om de tabel kort te houden.

Nu wordt met $\min_{(r,s) \in Y_{n-1} \times Y_n} f_{n,(r,s)}$ het 'lengte twee'-probleem tot een minimum beperkt. Laten we nog eens kijken naar het bovenstaande voorbeeld van San Jose. Er zijn vijf geografische verwijzingen: 'San Juan in Puerto Rico', 'San Jose in Costa Rica', 'San Jose, Costa Rica', 'San Jose, California' en 'San Jose'. Alleen de laatste verwijzing is ambigu. Stel dat we alleen San Jose, Californië en San Jose, Costa Rica als mogelijke alternatieven beschouwen. Als we T variëren, levert dit de volgende resultaten op:

$T = 1$	San Jose, Californië,
$T = 2$	San Jose, Californië *of*
	San Jose, Costa Rica,
$T = 3$	San Jose, Costa Rica,
$T = 4$	San Jose, Costa Rica.

Met een gematigde T wordt het algoritme minder gevoelig voor uitschieters.

Het algoritme levert vaak bevredigende resultaten op, maar heeft toch enkele zwakke punten. Ten eerste ligt aan het optimalisatieprobleem de veronderstelling ten grondslag dat de geografische verwijzingen in de tekst ertoe neigen om kleine clusters te vormen. Bovendien wordt ervan uitgegaan dat opeenvolgende locaties geografisch dicht bij elkaar zouden moeten liggen. Deze veronderstellingen houden mogelijk niet altijd stand. Ten tweede is geodetische afstand slechts een proxy voor de kans dat twee locaties vlakbij worden genoemd. Neem de volgende zin: "Ik ben net van New York naar Londen gevlogen."

Het is duidelijk dat de auteur van New York, New York naar Londen, Engeland is gevlogen en niet van New York, New York naar Londen, Ohio of van New York, Lincolnshire naar Londen, Engeland, terwijl beide bestemmingen geografisch dichterbij liggen. Voor de oplossing van dit probleem hebben we een betere afstandsfunctie tussen locaties nodig. Ook moeten we mogelijk aanvullende context opnemen in het algoritme. Het algemenere probleem bij het afleiden van gestructureerde gegevens uit ongestructureerde bronnen wordt gegevensextractie genoemd [7, 34]. Wat we in deze paragraaf doen, is een speciaal geval van extractie van geografische gegevens. In de conclusie stellen we een uitbreiding van ons algoritme voor disambiguering van benoemde entiteiten voor.

Metagegevens en titelremediatie

Zoals we in de inleiding hebben opgemerkt, beschikken we niet over tekst voor veel items waarmee we willen werken. Het is ook mogelijk om informatie uit de metagegevens zelf te proberen op te halen. De metagegevens bevatten vaak dubbelzinnige verwijzingen naar plaatsnamen of datums. Deze zouden we ter vervanging daarvan kunnen gebruiken. Het gebruik van metagegevens om de kwaliteit van deze gegevens te verbeteren wordt remediatie van metagegevens genoemd [8].

We bespreken eerst hoe het datumveld van een verzameling metagegevens kan worden geremedieerd. Het datumveld is met name belangrijk omdat mensen vaak naar items willen bladeren op basis van de tijdelijke relevantie. (Hoe vaak hebt u niet een e-mail opgezocht die u ongeveer twee maanden geleden hebt verstuurd?)

Het lijkt misschien een vreemd idee om metagegevens met zichzelf te remediëren. Per slot van rekening zijn metagegevens bedoeld om gestructureerde informatie over een artefact te verstrekken. Hoe kunnen we dit in hemelsnaam verbeteren? Dit is inderdaad mogelijk omdat de metagegevens mogelijk inconsistent zijn ingevoerd. Laten we een voorbeeld geven. Voor de collectie *gmd* in American Memory hebben we alle onderdelen van het MARC-veld onderzocht die de datumgegevens zouden moeten bevatten, bijvoorbeeld 260$c (publicatiedatum). In Tabel 5 wordt een overzicht weergegeven. Deze invoeritems zijn — als zodanig — enorm inconsistent en ongeschikt om een lijst met relevante items voor een bepaald jaar of een aantal jaren weer te geven. Om deze invoeritems te corrigeren, hebben we een ad-hocoplossing gekozen. In elk patroon dat we hebben aangetroffen, worden de jaargegevens vrijwel altijd aangegeven met de string ####. Omdat we dus de metagegevens wilden standaardiseren, converteerden we deze jaren naar een standaarddatumindeling en voerden we de gecorrigeerde metagegevens uit. We hoeven niet altijd ingewikkelde computertools te gebruiken.

Er is nog een uitdaging waarmee de Library of Congress wordt geconfronteerd bij veel van deze collecties: de metagegevens moeten in de loop der tijd worden verfijnd. Tijdens de eerste digitalisering van de *papr*-collectie van vroege speelfilms werd alleen een breedvoerige samenvatting van elke video verzameld. Zie Figuur 6 voor een voorbeeld. Op de meeste moderne websites, zoals YouTube, is vaak een korte titel voor elk item vereist. Deze titels moeten pakkend zijn en kunnen worden opgezocht om meer mensen te interesseren. Helaas waren de bestaande beschrijvingen te lang om als titel te fungeren. Omdat deze collectie minder dan duizend video's bevatte, kortte de Library of Congress handmatig elke beschrijving in tot een titel. We vroegen ons het volgende af: kunnen de beschrijvingen automatisch worden ingekort om een goede titel te verkrijgen? Nogmaals, zie de afbeelding voor een voorbeeld van onze titel van dezelfde video, vergeleken met de titel van de Library of Congress. In de gegenereerde titel wordt de essentie van de video beknopt vastgelegd. We bespreken in de volgende paragraaf hoe we onze gegenereerde titels hebben geëvalueerd, want hierbij deden zich enkele andere kwesties voor die we nader willen toelichten.

Titelsjablonen

Dan beschrijven we nu hoe we de titels genereren. Als eerste stap in het proces identificeren we gemeenschappelijke woordsoortpatronen in een bestaande database met titels. Deze patronen hebben de volgende vorm:

Samenvatting
Toont politieagenten en mannen met een hoge hoed en formele rijkleding terwijl ze grote bossen bloemen dragen tijdens een parade te paard. Wanneer de camerahoek enigszins verandert, verschijnt een marcherende band met een trommel waarop Bugle Corps, Lowell staat, gevolgd door een aantal gewapende militairen in uniform die in formatie marcheren. De camerahoek verandert zodat rijtuigen en de rest van de stoet in beeld worden gebracht. Het tafereel verplaatst naar een gebouw: de camera schuift langs de trap en toont een geestelijke in een lang gewaad die de kerk verlaat en teruggroet met zijn hoed in de hand. Geen titels.

Handmatige titel
St. Patrick's Day parade, Lowell, MA.

Onze titel
Parade van mannen te paard.

Figuur 6 Een voorbeeld van de manier waarop we automatisch de titel van een film genereren die in 1905 van een parade is gemaakt door Thomas Edison. De film is nu te zien op YouTube: http://www.youtube.com/watch?v=mKzcjKDgxHY.

Excavating	for	a	New	York	foundation
VBG	IN	DT	NNP	NNP	NN

Daarbij staan de codes voor respectievelijk: werkwoord, voorzetsel, lidwoord, eigennaam, eigennaam en zelfstandig naamwoord. We hebben deze berekend met de woordsoorttagger van Stanford [43]. Het idee is dat een grote titelverzameling gemeenschappelijke patronen in woordsoortreeksen zal bevatten. We kunnen de meest voorkomende patronen identificeren en als titelsjabloon gebruiken. Vervolgens kunnen we tekst uit de beschrijving afstemmen op de titelsjablonen en hopen dat het resultaat uit nuttige titels bestaat. Als eerste stap bij het genereren van titels moeten we dus een reeks titelsjablonen berekenen. Wer gebruikten de Newswire-collectie voor deze taak. Deze collectie bevat $1{,}3$ miljoen artikelen. Voor de titel van elk artikel berekenden we de woordsoortreeks en analyseerden we de patronen. Het resultaat is een database van 225.000 titelsjablonen.

Scores toewijzen aan woordgroepen

Voor het opbouwen van betekenisvolle titels moeten we betekenisvolle woordgroepen uit de beschrijving halen. We gebruiken een idee van Tomokiyo en Hurst [42]. Daarbij worden titels gezocht door een score toe te kennen aan een woordreeks op basis van twee maateenheden: de informatiewaarde en de woordgroepcohesie. Een reeks heeft een hoge informatiewaarde als de kans erg klein is dat deze reeks voorkomt in normale tekst. Een voorbeeld is de beschrijving 'singular value decomposition'. Het is zeer onwaarschijnlijk dat deze woordreeks voorkomt in een dagelijkse tekst, waardoor deze woordgroep zeer informatief is. De kans is echter aanzienlijk dat 'singular value decomposition' voorkomt in artikelen in het *SIAM Journal of Matrix Analysis*. De informatiewaarde van deze woordgroep staat dus in verhouding tot een achtergrondverzameling van 'standaardtekst'. Een woordgroep heeft een grote woordcohesie als de statistische eigenschappen van de woordgroep drastisch veranderen wanneer we de woordgroep opsplitsen. De woordgroep 'New York' heeft een hoge cohesie omdat in een document over 'New York' de woorden 'New' en 'York' vrijwel altijd samen zullen voorkomen. De statistische gegevens van 'New' en 'York' worden gekoppeld in dit document.

Deze concepten zijn geformaliseerd met een op n-grammen gebaseerd taalmodel tegen een achtergrondverzameling van tekst. Stel dat $\mathcal{C}$ een verzameling documenten is die als standaard worden beschouwd. De keuze van $\mathcal{C}$ is bepalend voor welke woorden als belangrijk worden gekozen in het bovenstaande voorbeeld met 'singular value decomposition', maar niet voor welke woorden als woordgroep worden beschouwd. Elke $d \in \mathcal{C}$ is in feite een reeks van woordtokens $d = (w_1, \ldots, w_m)$. Een op unigrammen gebaseerd taalmodel is de kans dat elk afzonderlijk woord voorkomt in de verzameling documenten. Een op bigrammen gebaseerd taalmodel is de kans dat elke woordenreeks voorkomt in de verzameling documenten.

Neem nu de woordenreeks in de beschrijving van een item: $d = (w_1, \ldots, w_m)$. Voor een woordenreeks (w_i, w_{i+1}, w_{i+2}) bedraagt de score voor de informatiewaarde:

$$P(w_i, w_{i+1}) = \text{Prob}\left[(w_i, w_{i+1}, w_{i+2}) \text{ in } d\right] \cdot \log\left(\frac{\text{Prob}\left[(w_i, w_{i+1}, w_{i+2}) \text{ in } d\right]}{\text{Prob}\left[(w_i, w_{i+1}) \text{ in } d\right] \cdot \text{Prob}\left[(w_{i+1}, w_{i+2}) \text{ in } d\right]}\right).$$

De score voor de informatiewaarde is:

$$I(w_i, w_{i+1}) = \text{Prob}\left[(w_i, w_{i+1}) \text{ in } \mathcal{C}\right] \cdot \log\left(\frac{\text{Prob}\left[(w_i, w_{i+1}) \text{ in } \mathcal{C}\right]}{\text{Prob}\left[w_i \text{ in } \mathcal{C}\right] \cdot \text{Prob}\left[w_{i+1} \text{ in } \mathcal{C}\right]}\right).$$

Deze scores zijn slechts de waarden voor de Kullback–Leibler-divergentie tussen het trigram- en bigrammodel in de beschrijving voor informatiewaarde en tussen het bigram- en unigrammodel in de achtergrondverzameling voor informatiewaarde. Met extreem korte beschrijvingen gebruiken we de achtergrondverzameling om de scores voor informatiewaarde te berekenen in plaats van de tekst van de beschrijving.

Een probleem met deze modellen is dat we gebeurtenissen met een kans van nul kunnen tegenkomen. Afgevlakte modellen zijn de standaardcorrectie voor deze kans van nul. Het idee achter een afgevlakt model is dat de kans op gebeurtenis niet nul is, zelfs niet als deze nog nooit is waargenomen. Een eenvoudig type afvlakking dat in statistiek wordt gebruikt, staat bekend als pseudo-count, waarvan Laplaciaanse afvlakking op basis van Laplace's regel van opeenvolging het klassieke voorbeeld is. Voor de kans dat een n-gram voorkomt in taal worden twee technieken veel gebruikt: Katz-afvlakking [25] en Kneser–Ney-afvlakking [27]. Bij Katz-afvlakking worden de gemeten aantallen verminderd met een vermenigvuldigingsfactor kleiner dan 1. De verwijderde aantallen worden gedistribueerd over de niet-waargenomen n-grammen op basis van het aantal n-grammen van een lagere orde, bijvoorbeeld het aantal unigrammen in plaats van het aantal bigrammen. Bij Kneser–Ney-afvlakking wordt additieve reductie in plaats van een vermenigvuldigingsfactor gebruikt. Ook kunnen hiermee beter n-grammodellen van een lagere orde worden opgebouwd waarin combinaties van meerdere woorden beter worden verwerkt. Stel dat 'San Francisco' veel voorkomt, maar dat 'Francisco' alleen voorkomt na 'San'. Kneser-Ney wijst aan 'Francisco' een lagere kans op een unigram toe omdat het woord alleen voorkomt in bepaalde bigram-combinaties, hetgeen tot uiting komt in hoge kansen op een bigram.

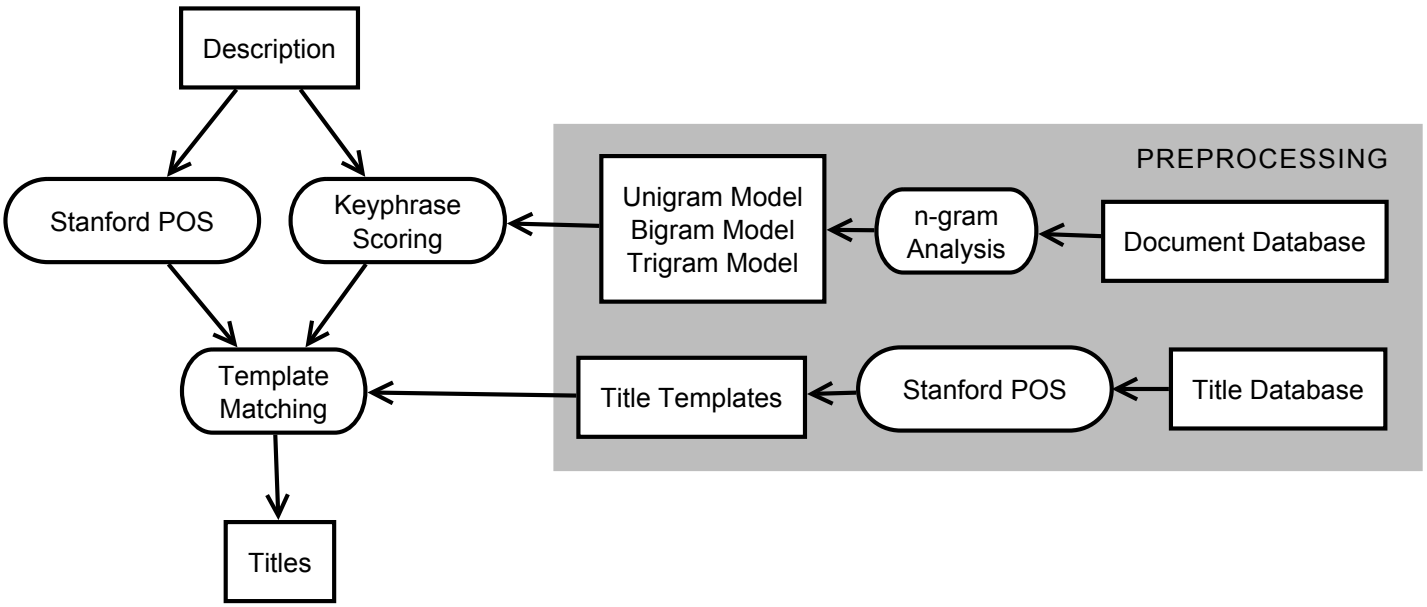
Samenvatting

Nadat we de scores van het nut van een bepaalde woordgroep hebben verkregen, hoeven we alleen nog maar de woordgroepen af te stemmen op de titelsjablonen om een titel te genereren. De titel met het hoogste gewicht (som van scores) is waarschijnlijk de beste titel.

Laten we het proces nog even samenvatten. Stel dat $\mathcal{C}$ een verzameling is van tekst met algemene informatie. Dit is de achtergrondverzameling. Bereken de kans op unigrammen, bigrammen en trigrammen in deze achtergrondverzameling. Stel vervolgens dat $\mathcal{T}$ een titelverzameling is. Bereken voor elke titel $t \in \mathcal{T}$ de woordsoortreeks voor de titel met behulp van de woordsoort-tool van Stanford (of een andere tool waarmee woordsoorten worden geïdentificeerd). Stel een reeks titelsjablonen samen op basis van deze woordsoortreeksen. Bereken nu op basis van een beschrijving de woordsoortreeks voor deze beschrijving. Bereken voor elke bigram in de beschrijving de score voor woordgroepcohesie en informatiewaarde. Neem de som van deze scores als totaalscore voor de sleutelwoordgroep van dit bigram. Stel vervolgens voor elke titelsjabloon een reeks bigrammen samen die overeenkomen met de woordsoortreeks. We vatten het proces samen in Figuur 7.

Conclusies en ideeën

Nog even onze motivering. Voor moderne verzamelingen van digitale gegevens zijn nieuwe zoektechnologieën nodig om deze gegevens relevant te maken zodat ze het waard zijn om te worden opgeslagen. Voor historische verzamelingen van gedigitaliseerde gegevens zijn geavanceerde zoekmethoden vereist zo-



Figuur 7 Ons proces voor het opbouwen van titels. Cirkels geven verwerking aan en rechthoeken geven data aan. Het grijze gebied geeft eenmalige voorverwerking aan. De rest van het proces moet voor elke beschrijving worden uitgevoerd.

dat mensen de artefacten kunnen vinden die ze interessant vinden. In beide scenario's zijn interessante metagegevens over de objecten van de digitale verzamelingen nodig. In dit artikel hebben we een algemene visualisatie van de onderwerptitels van de Library of Congress gepresenteerd. Dankzij deze visualisatie kregen we snel inzicht in een nieuwe verzameling gekoppelde gegevens. Op basis van onze ervaring met deze gegevensset onderzochten we een algoritme waarmee we de onderwerptitels in de collectie van de Library of Congress konden *afstemmen* op de categorieën van de Wikipedia-encyclopedie. Ons algoritme, dat is beschreven in [2], leverde bijna-optimale theoretische resultaten op; en een potentieel nuttige reeks overeenkomsten tussen de twee.

Vervolgens hebben we gekeken naar de desambiguering van geografische verwijzingen in een tekst. Dit leidde tot een eenvoudig, dynamisch programma waarbij we gebruikmaakten van de afstanden tussen mogelijke locaties die we eenvoudig kunnen oplossen. Voor de oplossing van geografische verwijzingen gaat het met name om het remediëren van metagegevens. We gingen verder met een ander onderzoek naar hoe we betere titels konden genereren op basis van korte beschrijvingen.

Met deze ideeën worden slechts enkele mogelijkheden onderzocht en onderzoek in digitaal beheer is in volle zwang, niet alleen bij ons, maar ook bij vele andere onderzoeksgroepen. Wij zijn bijvoorbeeld nu bezig met het verbeteren van geografische desambiguering, waar we niet alleen van locatieverwijzingen gebruik maken, maar ook gerelateerde informatie die we vinden in Wikipedia, of andere informatiebestanden. In de toekomst leggen we ons voornamelijk toe op meertalig zoeken en ontdekken [5, 9, 16]. In dit artikel hebben we nog niet gesproken over evaluatie van onze werkwijzen. In ons onderzoek maken we zelf vaak gebruik van Mechanical Turk van Amazon. Expliciete feedback van gebruikers kan ook worden benut en dit is een tweede richting die we hopen in te slaan in de nabije toekomst [33].

Gegevensbronnen

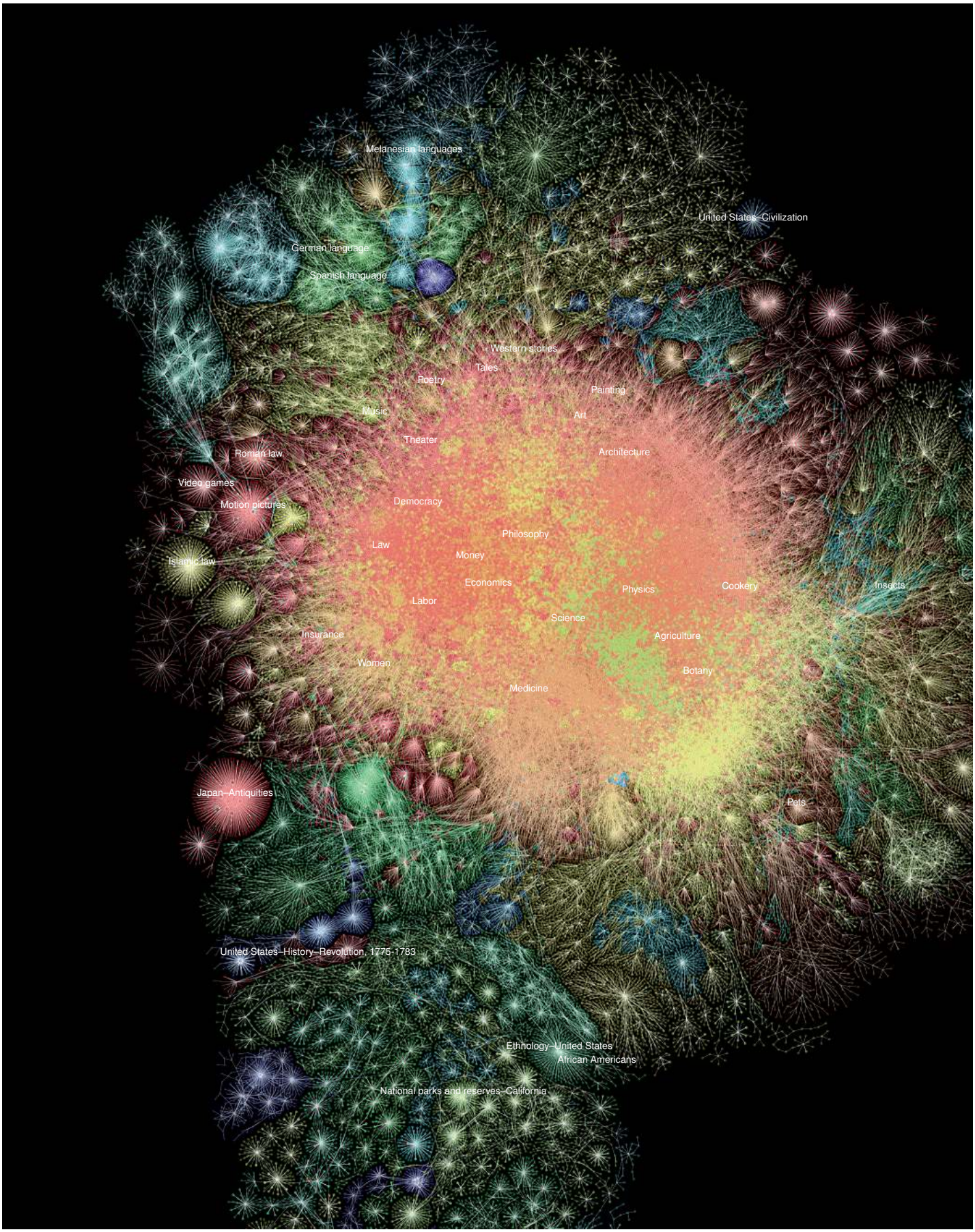
Een groot deel van dit artikel was gericht op hoe we met openbare gegevens de zoekervaring in de eigen gegevens van de Library of Congress kunnen verbeteren. Daardoor vragen lezers zich mogelijk af hoe ze een bijdrage kunnen leveren. Zoals we eerder hebben opgemerkt, heeft het succes van openbare gegevens geleid tot een overvloed aan vrij beschikbare gegevenssets. Dit zijn enkele van onze favorieten:

- Onderwerptitels van de Library of Congress: nu vrij beschikbaar, http://id.loc.gov/authorities.
- Rameau: de onderwerptitels van de Franse nationale bibliotheek, http://www.cs.vu.nl/STITCH/rameau/dump.
- Freebase: een grote verzameling gestructureerde en semigestructureerde informatie, http://freebase.com.
- Open Library: metagegevens over boeken, http://openlibrary.org.

Elke gegevensset biedt de gegevens bulksgewijs. Daardoor wordt het eenvoudig om de gegevens te interpreteren. Freebase bestaat uit allerlei kleine verzamelingen. Het ontwerpen van zoek- en bladertechnieken voor deze individuele verzamelingen lijkt enigszins op de eigen metagegevens van de Library of Congress.

Daarnaast willen we nog een mogelijkheid aanbevelen: neem contact op met verantwoordelijken op uw universiteit of in de nationale bibliotheek. We hebben de ervaring dat deze instellingen openstaan voor nieuwe ideeën en benaderingen. Als u deze weg volgt, moet u echter het geduld opbrengen om meer te leren over bibliotheek- en informatiewetenschappen (de historische thuisbasis voor het bestuderen van gegevensorganisatie en -toegang).

Dankwoord
We zijn zeer veel dank verschuldigd aan de behulpzame mensen van de Library of Congress, de Digitale wereldbibliotheek en het National Digital Information Infrastructure Preservation Program omdat ze ons hebben verteld over hun problemen bij het beheer van digitale gegevens en ons geduldig hebben geholpen tijdens onze voorlichting over hun problemen. Speciale dank gaat uit naar Laura Campbell, George Coulbourne, Beth Dulabahn, Jane Mandelbaum, Barbara Tillett en de bibliothecaris van het Congres, James Billington. Ook gaat onze dank uit naar Mohsen Bayati, die ons heeft geholpen bij het opstellen van een schaalbaar algoritme voor het netwerkalignmentprobleem, en Les Fletcher, die ons vaak heeft voorzien van nuttige adviezen.



Figuur 8 Deze tekening toont de grootste verbonden component van de ongerichte graaf met koppelingen in de onderwerptitels van de Library of Congress, waarbij knooppunten zijn gekleurd op basis van een cluster uit het CLUTO-programma. Enkele knooppuntlabels worden weergegeven om de onderwerpen in een specifieke regio te laten zien.

Referenties

1 A.T. Adai, S.V. Date, S. Wieland en E.M. Marcotte, LGL: creating a map of protein function with an algorithm for visualizing very large biological networks, *J. Mol. Biol.* 340(1) (2004), 179–190.

2 M. Bayati, M. Gerritsen, D.F. Gleich, A. Saberi en Y. Wang, Algorithms for large, sparse network alignment problems, in *Proceedings of the 9th IEEE International Conference on Data Mining*, 2009, pp. 705–710.

3 V.D. Blondel, A. Gajardo, M. Heymans, P. Senellart en P.V. Dooren, A measure of similarity between graph vertices: Applications to synonym extraction and web searching, *SIAM Review* 46(4) (2004), 647–666.

4 L.M. Chan, Still robust at 100: A centry of LC Subject Headings, in *Library of Congress Information Bulletin*, 1998.

5 P.A. Chew, B.W. Bader, T.G. Kolda en A. Abdelali, Cross-language information retrieval using PARAFAC2, in *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2007, pp. 143–152.

6 D. Conte, P.P. Foggia, C. Sansone en M. Vento, Thirty years of graph matching in pattern recognition, *International Journal of Pattern Recognition and Artificial Intelligence* 18(3) (2004), 265–298.

7 H. Cunningham, D. Maynard, K. Bontcheva en V. Tablan, GATE: A framework and graphical development environment for robust nlp tools and applications, in *Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics*, 2002.

8 G. de Groat, Future directions in metadata remediation for metadata aggregators, Technical Report, Digital Library Federation, 2009.

9 S.C. Deerwester, S.T. Dumais, T.K. Landauer, G.W. Furnas en R.A. Harshman, Indexing by latent semantic analysis, *Journal of the American Society of Information Science* 41(6) (1990), 391–407.

10 M. Ehrig en S. Staab, QOM – quick ontology mapping, in *Third International Semantic Web Conference*, 2004, pp. 683–697.

11 J.R. Finkel, T. Grenager en C. Manning, Incorporating non-local information into information extraction systems by gibbs sampling, in *ACL'05: Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, 2005, pp. 363–370.

12 C. Fraikin, Y. Nesterov en P.V. Dooren, A gradient-type algorithm optimizing the coupling between matrices, *Linear Algebra and its Applications* 429(5–6) (2008), 1229–1242. Special Issue devoted to selected papers presented at the *13th Conference of the International Linear Algebra Society*.

13 C. Fraikin, Y. Nesterov en P. Van Dooren, Optimizing the coupling between two isometric projections of matrices, *SIAM J. Matrix Anal. Appl.* 30(1) (2008), 324–345.

14 F. Göbel en A.A. Jagers, Random walks on graphs, *Stochastic Processes and their Applications* 2(4) (1974), 311–336.

15 A. Halevy, P. Norvig en F. Pereira, The unreasonable effectiveness of data, *IEEE Intelligent Systems* 24(2) (2009), 8–12.

16 R.A. Harshman, PARAFAC2: Mathematical and technical notes, *UCLA Working Papers in Phonetics* 22 (1972), 30–44.

17 J. Heer en M. Bostock, Crowdsourcing graphical perception: using mechanical turk to assess visualization design, in *CHI'10: Proceedings of the 28th international conference on Human factors in computing systems*, 2010, pp. 203–212.

18 N. J. Higham, *Handbook of Writing for the Mathematical Sciences*, SIAM, 1998.

19 W. Hu, Y. Qu en G. Cheng, Matching large ontologies: A divide-and-conquer approach, *Data Knowl. Eng.* 67(1) (2008), 140–160.

20 B.A. Huberman, D.M. Romero en F. Wu, Social networks that matter: Twitter under the microscope, *First Monday* 14(1) (2008).

21 A. Java, Twitter social network analysis, *UMBC ebquity blog*, 2007, http://ebiquity.umbc.edu/blogger / 2007 / 04 / 19 / twitter-social-network-analysis.

22 A. Java, X. Song, T. Finin en B. Tseng, Why we Twitter: understanding microblogging usage and communities, in *WebKDD/SNA-KDD'07: Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis*, 2007, pp. 56–65.

23 Y. Jia, J. Hoberock, M. Garland en J. Hart, On the visualization of social and other scale-free networks, *IEEE Transactions on Visualization and Computer Graphics* 41(6) (2008), 1285–1292.

24 G. Karypis, CLUTO – a clustering toolkit, Technical Report 02-017, University of Minnesota, 2002.

25 S.M. Katz, Estimation of probabilities from sparse data for the language model component of a speech recognizer, *IEEE Transactions on Acoustics, Speech and Signal Processing* 35(3) (1987), 400–401.

26 A. Kittur, E.H. Chi en B. Suh, Crowdsourcing user studies with mechanical turk, in *CHI'08: Proceeding of the twenty-sixth annual SIGCHI conference on Human factors in computing systems*, 2008, pp. 453–456.

27 R. Kneser en H. Ney, Improved backing-off for m-gram language modeling, in *International Conference on Acoustics, Speech, and Signal Processing, 1995. ICASSP-95*, 1995, pp. 181–184.

28 C. Körner, D. Benz, A. Hotho, M. Strohmaier en G. Stumme, Stop thinking, start tagging: tag semantics emerge from collaborative verbosity, in *WWW'10: Proceedings of the 19th international conference on World wide web*, 2010, pp. 521–530.

29 B. Krishnamurthy, P. Gill en M. Arlitt, A few chirps about Twitter, in *WOSP'08: Proceedings of the first workshop on Online social networks*, 2008, pp. 19–24.

30 T. Kuny, A digital dark ages? challenges in the preservation of electronic information, in *63rd International Federation of Library Associations and Institutions Council and General Conference (IFLA1997)*, 1997.

31 H. Kwak, C. Lee, H. Park en S. Moon, What is Twitter, a social network or a news media?, in *WWW'10: Proceedings of the 19th international conference on World wide web*, 2010, pp. 591–600.

32 A.Y. Levy, A. Rajaraman en J.J. Ordille, Querying heterogeneous information sources using source descriptions, in *VLDB'96: Proceedings of the 22th International Conference on Very Large Data Bases*, 1996, pp. 251–262.

33 S. Levy, How google's algorithm rules the web, *Wired Magazine* 18(3), 2010.

34 A. McCallum, Information extraction: Distilling structured data from unstructured text, *Queue* 3(9) (2005), 48–57.

35 S. Melnik, H. Garcia-Molina en E. Rahm, Similarity flooding: A versatile graph matching algorithm and its application to schema matching, in *Proceedings of the 18th International Conference on Data Engineering*, 2002, p. 117.

36 L. Ninove, *Dominant Vectors of Nonnegative Matrices: Application to Information Extraction in Large Graphs*, Ph.D. thesis, Université Catholique de Louvain, 2008.

37 D. Rafiei en S. Curial, Effectively visualizing large networks through sampling, *Visualization Conference, IEEE*, 2005, p. 48.

38 A. Rajaraman, More data usually beats better algorithms, *Datawocky Blog*, 2008, http://anand.typepad.com/datawocky/2008/03/more-data-usual.html.

39 J. Ross, L. Irani, M.S. Silberman, A. Zaldivar en B. Tomlinson, Who are the crowdworkers?: shifting demographics in mechanical turk, in *CHI EA'10: Proceedings of the 28th of the international conference extended abstracts on Human factors in computing systems*, 2010, pp. 2863–2872.

40 M.S. Silberman, J. Ross, L. Irani en B. Tomlinson, Sellers' problems in human computation markets, in *HCOMP'10: Proceedings of the ACM SIGKDD Workshop on Human Computation*, 2010, pp. 18–21.

41 J. Surowiecki, *The Wisdom of Crowds: Why the Many Are Smarter Than the Few and How Collective Wisdom Shapes Business, Economies, Societies and Nations*, Little, Brown, 2004.

42 T. Tomokiyo en M. Hurst, A language model approach to keyphrase extraction, in *Proceedings of the ACL 2003 workshop on Multiword expressions*, 2003, pp. 33–40.

43 K. Toutanova, D. Klein, C.D. Manning en Y. Singer, Feature-rich part-of-speech tagging with a cyclic dependency network, in *NAACL'03: Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, 2003, pp. 173–180.

44 Various, The MARC Standard, 2007, http://www.loc.gov/marc, accessed on 17 September 2007.

45 J. Weng, E.-P. Lim, J. Jiang en Q. He, TwitterRank: finding topic-sensitive influential twitterers, in *WSDM'10: Proceedings of the third ACM international conference on Web search and data mining*, 2010, pp. 261–270.

46 K.-P. Yee, K. Swearingen, K. Li en M. Hearst, Faceted metadata for image search and browsing, in *CHI'03: Proceedings of the SIGCHI conference on Human factors in computing systems*, 2003, pp. 401–408.

47 E. Yeh, D. Ramage, C.D. Manning, E. Agirre en A. Soroa, Wikiwalk: random walks on wikipedia for semantic relatedness, in *TextGraphs-4: Proceedings of the 2009 Workshop on Graph-based Methods for Natural Language Processing*, 2009, pp. 41–49.