



HOME ADVERTEREN ABONNEREN REDACTIE CONTACT **SITMAP**

LOGIN

ZOEKEN

GEAVANCEERD

NIEUWSOVERZICHT

AGENDA

LINKS

ARCHIEF

VAKPUBLICATIES

IP LEZING

VACATURES

BINNENKORT

WEBTESTEN

OVER DEZE SITE

RSS FEEDS

IP FLASH

WEBLOG

LEVERANCIERSGIDS



Met spoed een media-analyse presenteren?

STITCH

Het webdossier Catch met vele extra's bij onderstaand artikel is te vinden op informatieprofessional.googlepages.com

Digitaal zoeken in meerdere collecties tegelijk

Volautomatisch overeenkomsten tussen verschillende thesauri vinden. Dat is informatici van de Vrije Universiteit samen met medewerkers van de Koninklijke Bibliotheek gelukt. Zo moet zowel de gewone bibliotheekgebruiker als de informatieprofessional in de toekomst met één zoekopdracht door meerdere bibliotheekcollecties kunnen zoeken, niet alleen in eigen land maar ook in het buitenland.

[Door: Bennie Mols]

De Koninklijke Bibliotheek (KB) in Den Haag beschikt over meer dan tachtig kilometer aan boeken en tijdschriften. Daaronder bevinden zich tientallen bijzondere collecties, zoals middeleeuwse handschriften, middeleeuwse illustraties, maar ook strips, kookboeken en affiches. Het is historisch zo gegroeid dat elke collectie vaak wordt beschreven met een eigen catalogus of thesaurus. De ene thesaurus gebruikt bijvoorbeeld de zoekterm 'plankzeilen', terwijl een andere de zoekterm 'surfsport' gebruikt.

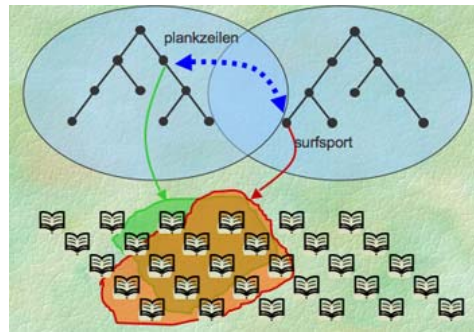
Nu bibliotheken hun catalogi via internet voor iedereen en van over de hele wereld toegankelijk hebben gemaakt, wordt het steeds urgenter om met dezelfde zoekterm in meerdere boekencollecties tegelijk te kunnen zoeken, en niet voor de ene collectie 'plankzeilen' te moeten gebruiken en voor de andere 'surfsport,' of de ene keer 'influenza' en de andere keer 'griep'.

Het zoeken in boekencollecties die met verschillende thesauri worden beschreven, wordt nog door een aantal andere factoren bemoeilijkt. Zo kunnen persoonsnamen ambigu zijn. Wil de gebruiker die de zoekterm 'van Gogh' intikt, iets weten over de schilder Vincent van Gogh of over de filmer Theo van Gogh, of misschien nog een andere Van Gogh? Verder wordt met verschillende benamingen soms hetzelfde bedoeld. Dezelfde Japanse kunststroming die in Nederland de 'Edo-periode' heet, wordt elders ook wel aan geduid met 'Tokugawa-periode'. En wie met een druk op de knop tegelijk wil zoeken in de KB en in de Franse zuster, de Bibliothèque Nationale de France (BNF), loopt zowel op tegen een taalprobleem als een probleem met verschillende thesauri.

Dubbele annotatie

In het onderzoeksproject STITCH (Semantic Interoperability to Access Cultural Heritage) – het Engelse 'to stitch' betekent 'aan elkaar rijgen' – werken onderzoekers van de Vrije Universiteit (VU) in Amsterdam samen met medewerkers van de KB aan twee projecten die verschillende boekencollecties tegelijk doorzoekbaar moeten maken. 'De kunst is om collecties te integreren alleen op basis van metadata,' zegt Frank van Harmelen, STITCH-projectleider en hoogleraar kunstmatige intelligentie aan de VU. 'Wij zoeken nooit in de boeken zelf. Immers, de inhoud van veruit de meeste boeken is nog niet digitaal beschikbaar, en het is maar de vraag of dat ooit het geval zal zijn. We kijken daarom alleen in de catalogi die de boeken beschrijven, want die zijn wel algemeen digitaal beschikbaar.' Zo heeft de KB een Wetenschappelijke Collectie van anderhalf miljoen boeken en een Depotcollectie van een miljoen boeken. Beide collecties worden met een aparte thesaurus beschreven: de Wetenschappelijke Collectie met een thesaurus van 35.000 termen en de Depotcollectie met een van 5.000 termen. Elke thesaurus is een hiërarchisch woordenboek dat begint met algemene termen, en naar beneden toe steeds specifiek wordt. Onder de zoekterm 'planten' hangt bijvoorbeeld de zoekterm 'bomen' en daaronder 'eiken,' 'beuken' enzovoort.

Van Harmelen: 'Dat betekent dat de gebruiker veroordeeld is tot tweemaal zoeken: met de ene thesaurus in de Wetenschappelijke Collectie en met de andere in de Depotcollectie. Daarnaast wordt een kwart miljoen boeken met beide thesauri beschreven. Dat betekent niet alleen dubbele annotatiekosten, maar ook dubbele onderhoudskosten. Zo werden in 2006 1700 boeken dubbel geannoteerd. In principe hebben we dat probleem nu opgelost. In de afgelopen 2,5 jaar hebben we samen met de KB een methode ontwikkeld om in één keer in beide catalogi te zoeken.'



Automatische statistische coördinatie

Statistische methode
De wetenschappelijke truc zit in een statistische methode die volautomatisch kijkt hoe sterk

twee verzamelingen overlappen, en die de mate van overlap in een getal uitdrukt. Hoe sterker de overlap, hoe groter het getal. De statistische methode weet zelf niets van betekenissen van woorden, maar kan wel snel overlappende verzamelingen opsporen. Omdat een kwart miljoen boeken met twee thesauri is beschreven, konden de onderzoekers kijken in welke mate zoektermen uit beide thesauri dezelfde boeken beschreven. Als zoeken op 'plankzeilen' vrijwel dezelfde boeken oplevert als zoeken op 'surfsport', dan tekent het

FOTO VAN DE WEEK



Het Scription in Tilburg staat tijdens het museumweekend op 5 en 6 april in het teken van Typewriter Day en de expositie 'The Singing Typewriter'.
> lees meer

MAGAZINE



> lees meer

CATCH-DOSSIER

InformatieProfessional besteedt in een reeks artikelen aandacht aan tien CATCH-projecten, waarin er goedgeerd en informatica samenkomen. In een webdossier zijn de bijdragen met extra bijlagen te lezen.

VACATURES

Koninklijk Instituut voor de Tropen (Amsterdam) zoekt een Applicatiebeheerder/ Functioneel ontwikkelaar bibliothecaire diensten

Nuffic (Den Haag) zoekt een Informatieanalist

BINNENKORT

Informatiebijeenkomst over subsidieregeling 'Digitaliseren met beleid'
> lees meer

Symposium Auteursrecht
> lees meer

LAATSTE REACTIES

"Onze Lage Landen kwamen in het verleden slechts beperkt aan hun trekken..."
> lees meer

Bedankt!
> lees meer

Gefeliciteerd! Ik volg je weblog en je weet me regelmatig te verrassen.
> lees meer

Edwin, gefeliciteerd! Je inspanningen worden nu ook 'landelijk' erkend,...
> lees meer

Ik ben benieuwd hoeveel meer mensen zich aangesproken voelen om archieven op...
> lees meer

hetzelfde onderwerp gaat. Omdat de ene thesaurus 35.000 termen bevat en de andere 5.000, hebben niet alle termen uit de een, ook automatisch een equivalent in de ander. Maar omdat de vocabulaires hiërarchisch zijn opgebouwd, weet het zoekprogramma wel dat als een term hoger in de boomstructuur een equivalent heeft in de ander, de kans groot is dat ook de takken daaronder met elkaar te maken hebben. Zo wordt de hiërarchie van het vocabulaire gebruikt om meer termen te overdekken dan het programma in de eerste plaats vindt. Afhankelijk van wat de eindgebruiker wil, kan hij instellen in welke mate correctheid en volledigheid van het gezochte belangrijk zijn. Een scholier die een scriptie schrijft, zal al blij zijn als hij 50 goede artikelen vindt van een collectie van 100, terwijl een wetenschapper ze alle 100 wil vinden, ook als hij daarvoor van de 150 zoekresultaten er zelf handmatig 50 moet wegstrepen.

Brinkman	GTT	
Diabetes mellitus	Suikerziekte	De mate van overlap, gerelateerd aan de wens van de gebruiker, bepaalt of termen gematcht worden
MER	milieueffect-rapportages	
Lyme ziekte	Ziekte van Lyme	
Plankzeilen	Surfsport	
Influenza	Griep	
Neus	Ademhalingsorganen	
Zonde	Genade	
Ketters	Inquisitie	

IJking van de methode

Hoe goed werkt de methode? Van Harmelen: 'Als we de parameters heel conservatief instellen, en alleen maar de 1000 beste antwoorden opvragen, dan blijkt daarvan 90 procent correct te zijn. Dat is op zich mooi, maar we missen dan natuurlijk heel veel antwoorden die ook goed zijn, maar die niet bij de eerste 1000 zitten. Als we de parameters wat toleranter instellen zodat we meer goede antwoorden terugkrijgen, bijvoorbeeld 70 procent, dan gaat dat ten koste van de precisie: niet alleen krijgen we dan meer goede antwoorden terug, maar helaas ook meer foute. Ruwweg is dan ongeveer 3 op de 10 antwoorden fout. Dus: 70 procent precisie op het 70 procent recall-niveau.'

Vervolgens zijn de resultaten van het zoekprogramma ter controle aan beroepsannotators van de KB gegeven. Zij beoordeelden of de woorden die volgens onze software hetzelfde betekenen ook echt hetzelfde betekenen. 'De KB-professionals vormen de gouden standaard voor het ijken van onze statistische methode,' aldus Van Harmelen. 'Juist door die nauwe samenwerking met de KB zijn we als een van de eersten in de wereld in staat geweest om onze statistische methode zo goed te gebruiken en zo uitgebreid te evalueren. Veel buitenlandse collega's publiceerden wel over hun wetenschappelijke methode, maar hadden geen mogelijkheid om die ook uitgebreid te testen op echte catalogi.'



KB
-
manuscript
illuminations

Gebruiksklaar
maken
Dit
deel
van
STITCH
is
inmiddels
afgerond.
De
KB
bekijkt

nu hoe de wetenschappelijke methode gebruiksklaar kan worden gemaakt. Van Harmelen: 'Wij hebben een nieuwe methode ontworpen en getest om in verschillende catalogi tegelijk te zoeken. Nu we hebben laten zien dat het onderliggende wetenschappelijke principe werkt, kan een commerciële partij er echt een product van gaan maken. Dat is niet meer onze taak, en dat begrijpt de KB prima.'

Er bestaat een groot aantal uiteenlopende technieken om verschillende catalogi te integreren: statistiek, logica, taalkunde en zelfs wiskundige grafentheorie. De grote vraag is nu wanneer welke methode het beste werkt. Van Harmelen en zijn onderzoekers werken nu verder aan het beantwoorden van deze overkoepelende vraag. 'Als morgen het Van Gogh Museum aankomt met de vraag om verschillende van hun catalogi te integreren, dan willen we een theorie uit de kast kunnen trekken die voorspelt welke methode we het beste kunnen gebruiken. Zo'n theorie bestaat nog niet en daar zoeken we nu naar.'



Iconclass
-
classificatie

Europese
boekenintegratie
De
STITCH
-
onderzoekers
werken
ook
nog
aan
een
tweede
project,
dat
binnen
een
groter,
Europees
kader

van de EU-landen te integreren. STITCH heeft inmiddels een pilotproject achter de rug dat zich richtte op het integreren van de catalogi die middeleeuwse illustraties van de Koninklijke Bibliotheek en de Bibliothèque Nationale de France beschrijven. Het probleem zit niet alleen in het gebruik van verschillende talen. Het zit vooral in een verschillende beschrijving van de wereld van de illustraties.

De BNF gebruikt echter haar eigen thesaurus, die Mandragore heet. Waar Iconclass bijvoorbeeld de zoekterm 'Religion and magic' gebruikt, doet de Franse Mandragore het met zoektermen als 'Christianisme', 'Autres religions' en 'Parapsychologies, occultisme, demonologie...'. STITCH zoekt naar automatische oplossingen om deze twee catalogi te integreren. Daarvoor hoeft geen nieuwe overkoepelende catalogus te worden gemaakt, want dat zou enorm veel extra handwerk vereisen. In dit geval kon geen statistische methode worden gebruikt omdat er geen objecten waren die in beide collecties voorkwamen. De Franse thesaurus werd nu eerst vertaald, deels met een automatisch woordenboek en deels met een al in het Frans bestaande subthesaurus van Iconclass, die een klein deel van Iconclass in het Frans beschrijft. Het algoritme dat na de vertaalslag zocht naar de overeenkomsten, gebruikte een combinatie van 'morfologische regels' en 'woordafstanden'. Morfologische regels beoordelen op grond van woordvormen (enkelvoud-meervoud; zelfstandig naamwoord; bijvoeglijk naamwoord, samengestelde woorden...) hoe sterk woorden op elkaar lijken. Woordafstanden bepalen een soort wiskundige afstand tussen woorden door te kijken hoeveel bewerkingen er nodig zijn om van het ene woord het andere te maken: hoeveel letters moet je toevoegen, weghalen of veranderen.

Mandragore - thesaurus
De KB gebruikt als thesaurus Iconclass, dat sinds de jaren vijftig in Nederland is ontwikkeld.



De automatische technieken brengen de integratie van internationale boekencollecties van nationale bibliotheken een flinke stap dichterbij. 'Met een automatische vergelijking van thesauri kun je op manieren zoeken die met gewone tekstzoekmachines niet kunnen,' besluit Van Harmelen. 'Bijvoorbeeld meertalig zoeken; het zoeken in collecties die niet uit tekst, maar uit illustraties bestaan; het integreren van collecties en ook het zoeken op basis van domeinkennis.' De binnen STITCH ontwikkelde technieken zijn zo algemeen dat ze ook buiten het domein van het culturele erfgoed bruikbaar zijn, bijvoorbeeld binnen de geneeskunde, waar vaak ook verschillende termen voor hetzelfde begrip worden gebruikt. Dat is het mooie van wiskunde en informatica: ze leveren gereedschap dat universeel toepasbaar is.

Bennie Mols is wetenschapsjournalist.

Wat is CATCH?

CATCH (Continuous Access To Cultural Heritage) is een nationaal onderzoeksprogramma van de NWO-gebieden exacte- en geesteswetenschappen. Binnen CATCH worden methoden en technieken ontwikkeld waarmee erfgoedbeheerders hun digitale collecties beter toegankelijk kunnen maken. Informaticaonderzoekers en erfgoedinstellingen zoals het Rijksmuseum, de KB, het Nationaal Archief, Naturalis en het Nederlands Instituut voor Beeld en Geluid werken in CATCH nauw samen. De onderzoekers verrichten het grootste deel van hun onderzoek dan ook binnen de instelling, die optreedt als projectbegeleider, eindgebruiker, en leverancier van data en content.

Lees ook:

- > de inleiding tot de serie Catch: [Onderzoeksprogramma brengt erfgoed en informatica samen](#)
- > [Hulptroepen bij de ontsluiting van av-materiaal](#)
- > [Digitaal zoeken in handgeschreven archieven](#)

Er is ook een [webdossier Catch](#) met vele extra's

Reacties

Er zijn nog geen reacties geplaatst.

Plaats een reactie

Naam:

plaats reactie

annuleer

[< terug](#)

 © 2002 - 2008 Infolook Nederland B.V. Alle rechten voorbehouden.