Hoe kan de ‘onderzoekende zoeker’ beter ondersteund worden?

Zoekmachines op internet zijn heel geschikt om allerlei soorten informatie te vinden over een onderwerp. Als ik in Google de zoekopdracht ‘Multatuli’ geef dan krijg ik 811.000 verwijzingen naar websites met informatie over de 19e-eeuwse Nederlandse schrijver. Maar als ik specifieke informatie over Multatuli zoek, zoals wat voor verbanden er zijn tussen zijn werk en politieke gebeurtenissen uit die tijd, of hoe andere schrijvers op zijn boeken reageerden, dan is de Google-strategie niet meer zo behulpzaam. Google geeft mij namelijk altijd verwijzingen naar complete webpagina’s, hoe specifiek mijn zoekvraag ook is.

Complexe zoekvragen door ‘onderzoekende zoekers’

Complexe zoekvragen, die bijvoorbeeld in gaan op verbanden tussen entiteiten, zijn voorbeelden van vragen die een historicus of literatuurwetenschapper zou kunnen hebben. Als deze onderzoeker geïnteresseerd is in literaire bronnen, dan kan hij of zij terecht bij gespecialiseerde tekstcollecties, bijvoorbeeld de Digitale Bibliotheek voor de Nederlandse Letteren [1]. DBNL heeft, net als de meeste andere online tekstcollecties, een zoekinterface die is geïnspireerd op Google: als ik de zoekvraag ‘Multatuli’ ingeef in het DBNL-zoeksysteem [2] dan krijg ik 127 pagina’s met resultaten. Als ik op één van de gevonden fragmenten klik, dan krijg ik een document te zien waarin alleen mijn zoekterm Multatuli is gemarkeerd. De tekst die ik op mijn scherm heb, is lang en ik zal hem moeten lezen om erachter te komen waar de informatie staat waar ik naar op zoek ben.

I. *Max Havelaar, of de koffij-veilingen der Nederlandsche Handel-Maatschappij, door* Multatuli. *Twee deelen. Tweede druk.* (*Te*) *Amsterdam*, (*bij*) J. de Ruijter, 1860. *Prijs* ƒ 4,00.
II. *Waarheen? Een woord aan de lezers van Max Havelaar. Tweede druk.* (*Te*) *Leijden*, (*bij*) P. Engels, 1860. *Prijs* ƒ 0,40.
In een tijdschrift dat den naam van Vaderlandsche Letteroefeningen draagt, mag een werk dat zooveel sensatie gemaakt heeft in den lande als *Max Havelaar*, niet onvermeld blijven, vooral wanneer zulk een werk een literarisch meesterstuk is. Dat eerst nu eene aankondiging van Multatuli’s geschrift plaats heeft, wijte men den uitgever de Ruijter, die geen exemplaar van het werk *ter recensie* zond aan de redactie van een tijdschrift, waarin alleen bij zeldzame uitzondering niet ingezonden boekwerken worden aangekondigd. Onze uitgever, de heer van der Mast, gaf een blijk van zijnen ijver door het kwaad te verhelpen, en zond mij eerst kort geleden *Max Havelaar* ter aankondiging. Mijne taak is door deze min wenschelijke vertraging zeer moeijelijk geworden. Reeds is er zoo menig woord over het bijna alom gelezen boek gevallen, zoo menig verschillend oordeel daarover geveld, dat ik bijna schroom onzen lezers mijnen schralen mosterd na den maaltijd voor te zetten. Vooral gevoel ik dien schroom, wanneer ik mij herinner, welk eene meesterlijke beschouwing van *Max Havelaar* Prof. P.J. Veth

‘Onderzoekende zoekers’ zoals de literatuurwetenschapper in dit voorbeeld zijn bereid tijd te besteden aan het bestuderen van tekstuele bronnen. Daarbij zouden ze echter veel beter geholpen kunnen worden dan met alleen het tonen van volledige documenten. Een eerste stap zou hierbij zijn als het zoeksysteem namen en gebeurtenissen

I. *Max Havelaar, of de koffij-veilingen der Nederlandsche Handel-Maatschappij, door* Multatuli. *Twee deelen. Tweede druk.* (*Te*) *Amsterdam*, (*bij*) J. de Ruijter, 1860. *Prijs* ƒ 4,00.
II. *Waarheen? Een woord aan de lezers van Max Havelaar. Tweede druk.* (*Te*) *Leijden*, (*bij*) P. Engels, 1860. *Prijs* ƒ 0,40.
In een tijdschrift dat den naam van Vaderlandsche Letteroefeningen draagt, mag een werk dat zooveel sensatie gemaakt heeft in den lande als *Max Havelaar*, niet onvermeld blijven, vooral wanneer zulk een werk een literarisch meesterstuk is. Dat eerst nu eene aankondiging van Multatuli’s geschrift plaats heeft, wijte men den uitgever de Ruijter, die geen exemplaar van het werk *ter recensie* zond aan de redactie van een tijdschrift, waarin alleen bij zeldzame uitzondering niet ingezonden boekwerken worden aangekondigd. Onze uitgever, de heer van der Mast, gaf een blijk van zijnen ijver door het kwaad te verhelpen, en zond mij eerst kort geleden *Max Havelaar* ter aankondiging. Mijne taak is door deze min wenschelijke vertraging zeer moeijelijk geworden. Reeds is er zoo menig woord over het bijna alom gelezen boek gevallen, zoo menig verschillend oordeel daarover geveld, dat ik bijna schroom onzen lezers mijnen schralen mosterd na den maaltijd voor te zetten. Vooral gevoel ik dien schroom, wanneer ik mij herinner, welk eene meesterlijke beschouwing van *Max Havelaar* Prof. P.J. Veth

zou markeren in het document dat de onderzoeker bekijkt. Als ik bijvoorbeeld een recensie van de Max Havelaar lees in DBNL zouden titels van boeken, schrijvers en tijdschriften (*Vaderlandsche Letteroefeningen*, bijvoorbeeld) gemarkeerd kunnen worden. Een volgende stap zou zijn om op verzoek van de gebruiker informatie te tonen over de gemarkeerde termen. Die informatie zou bijvoorbeeld uit een encyclopedie kunnen komen. Welke informatie precies van belang is, hangt af van de doelgroep van het zoeksysteem. Als ik op ‘Vaderlandsche Letteroefeningen’ klik, zou ik bijvoorbeeld informatie uit een dossier van de Koninklijke Bibliotheek kunnen krijgen: “*Vaderlandsche Letteroefeningen was meer dan een eeuw lang een van de toonaangevende literair-culturele tijdschriften van Nederland.*”, met een link naar externe bronnen en andere plaatsen in DBNL waar het tijdschrift genoemd wordt. Als documenten in tekstcollecties zoals DBNL verrijkt worden met dit soort informatie,

kunnen ‘onderzoekende zoekers’ zoals literatuurwetenschappers en historici in de toekomst beter geholpen worden bij hun literatuurstudie.

Taaltechnologie die onderzoekend zoeken mogelijk maakt

Het markeren van belangrijke termen en namen in tekst is een taak die al goed mogelijk is met bestaande taaltechnologie. Zo worden in PoliticalMashup [3], een samenwerkingsproject van NRC Handelsblad en het Informatica Instituut van de Universiteit van Amsterdam, politieke debatten automatisch geanalyseerd op onderwerpen die aan bod komen per partij en politicus, en welke namen en termen daarbij het vaakst gebruikt worden.

Voor het vergaren van feitelijke informatie die gekoppeld kan worden aan termen t.b.v. het ondersteunen van zoekers is nog meer onderzoek nodig. Daarvoor moet namelijk niet alleen een naam herkend worden als een entiteit (Multatuli is Eduard Douwes Dekker), maar ook de relatie met andere entiteiten gevonden worden (“Multatuli schreef de Max Havelaar”). Die relaties kunnen op veel verschillende manieren worden uitgedrukt: “de Max Havelaar is door Multatuli geschreven”, “de auteur van de Max Havelaar is Multatuli”, etc.

Initiatieven wereldwijd

In het onderzoeksproject *Extracting factoids from Dutch texts*, dat in 2011 wordt uitgevoerd aan de Radboud Universiteit Nijmegen, wordt hieraan gewerkt. Met financiering van Google in het kader van het Europese *digital humanities*-programma wordt in dit project software ontwikkeld die feitelijke informatie vergaart uit Nederlandse Wikipedia-teksten. Met behulp van die software kunnen in de toekomst tekstuele bronnen verrijkt worden met achtergrondinformatie. Google ondersteunt met de *digital humanities awards* de ontwikkeling van technologie ten behoeve van onderzoekers in de geesteswetenschappen [4]. In de Verenigde Staten worden momenteel een aantal grote projecten uitgevoerd die zich bezighouden met het automatisch leren van feiten uit teksten [5]. Sommige wetenschappers vinden dat zoeksystemen op internet veel preciezer te werk zouden moeten gaan in het beantwoorden van vragen. In augustus van dit jaar verscheen daarover zelfs een artikel in Nature, geschreven door de Oren Etzioni, de directeur van het Turing Center van de universiteit van Washington in Seattle [6]. Hij noemt onder andere de doorbraak van de IBM-computer Watson, die een kennisquiz won van de beste menselijke deelnemers. Op http://tiny.cc/jeopardy kun je zien hoe Watson de quiz ‘Jeopardy’ won. IBM onderzoekt nu de mogelijkheden om Watson in te zetten voor zoekvragen in plaats van trivia-vragen.

1) DBNL: www.dbnl.org
2) www.dbnl.org/zoeken/zoekeninteksten
3) www.nrc.nl/denhaag
4) Alle projecten die dit jaar worden gefinancierd staan op tiny.cc/googleawards
5) rtw.ml.cmu.edu/rtw
6) www.nature.com/nature/journal/v476/n7358/full/476025a.html

Auteur: Suzan Verberne, Centre for Language and Speech Technology, Radboud Universiteit Nijmegen