ZOEKMACHINES: SAMEN EN DUURZAAM VOORUIT

Rede uitgesproken bij de aanvaarding van het ambt van hoogleraar federatieve zoekmachines aan de Faculteit der Natuurwetenschappen, Wiskunde en Informatica van de Radboud Universiteit op vrijdag 1 maart 2024

door prof. dr. ir. Djoerd Hiemstra

Djoerd Hiemstra pleit in zijn inaugurele rede voor meer samenwerking en duurzaamheid bij het maken van zoekmachines. In de rede deelt hij een oude wijsheid over samenwerking; hij bespreekt zijn plan om studenten met verschillende achtergronden hun gedeelde geschiedenis te leren; en hij onthult zijn droom om onbeperkte toegang tot alle online informatie mogelijk te maken, met behulp van zoekmachines die *samen* worden ontwikkeld en onderhouden. De rede bevat auto's, iPhone opladers, de Space Shuttle en verwijzingen naar opwindend recent onderzoek.

Djoerd promoveerde in 2001 aan de Universiteit Twente op een proefschrift over het gebruik van taalmodellen voor het vinden van informatie. Djoerd werkte daarna aan verschillende *open source* zoekmachines: PuppyIR, een zoekmachine voor kinderen; PF/Tijah, een zoekmachine voor gestructureerde gegevens; MIREX, een zoekmachine voor experimenten met heel grote collecties; en Searsia, een federatieve zoekmachine. Sinds 2019 werk hij aan de Radboud Universiteit als hoogleraar met de leerstoel federatieve zoekmachines. Hij is verbonden aan het institute voor Computing en Information Sciences (iCIS) aan de faculteit der Natuur-wetenschappen, Wiskunde en Informatica.

Mijnheer de rector magnificus, geachte aanwezigen,

Welkom aan de Radboud Universiteit. Het is mij een grote eer om deze rede, een rede over zoek-machines, te mogen uitspreken, maar voordat ik begin, wil ik graag een oude wijsheid met u delen.

DE SIKSIKA

Het gaat om een oude wijsheid van de Siksika, een volk uit Noord Amerika dat door de witte amerikanen “Blackfoot Indians”, ofwel Zwartvoetindianen, wordt genoemd. De Siksika leven in de staat Montana in de Verenigde Staten en de staat Alberta in Canada. Nu moet u weten dat het daar ‘s winters erg koud kan worden. Zo koud, dat één deken onvoldoende is om iemand de hele nacht warm te houden. Als één deken onvoldoende is, wat doe je dan? Zorgen dat je twee dekens hebt? Dekens verkopen in de winter? Een start-up beginnen die alle dekens in de wijde omgeving opkoopt, zodat je de dekens heel duur kunt verkopen?[1] Nee, de oude wijsheid zegt het volgende: *Één persoon onder een deken vriest ‘s nachts dood, maar twee personen, onder precies zo’n zelfde deken, houden elkaar warm en overleven de koude winternacht.*

SAMENWERKEN

Dit is de essentie van samenwerken: Samenwerken maakt het resultaat beter (je leeft nog na een koude winternacht) én je hebt minder hulpmiddelen nodig (in plaats van één deken per persoon, gemiddeld maar een halve deken per persoon). Samen en duurzaam. Om samenwerking mogelijk te maken, moeten mensen bereid zijn om te delen. Je moet bereid zijn om je deken te delen met een ander gedurende een koude winternacht. Helaas, is het tegenwoordig in Nederland zo dat veel landgenoten hun deken liever niet delen met een ander, als die ander een vluchteling is, of als die ander een huidskleur, levensovertuiging of religie heeft die anders is dan die van hunzelf. Een deel van de Nederlanders vriest nog liever dood, dan dat ze hun deken delen met iemand uit Marokko.[2] Hier zie ik een belangrijke taak voor ons onderwijs. Ons onderwijs kan beter worden ingezet om begrip te kweken voor de ander. Begrip voor de ander is een voorwaarde voor samenwerking.

ONDERWIJS EN SAMENWERKEN

Mijn doel aan de Radboud Universiteit is om *studenten met verschillende achtergronden hun gedeelde geschiedenis te leren*. Dit is belangrijk voor het onderwijs op alle niveaus (basisschool, middelbare school, beroepsonderwijs en de universiteit) en voor alle vakgebieden (geschiedenis, wiskunde, psychologie, autotechniek, en ook mijn vakgebied de informatica). Wij delen onze geschiedenis, ook onze wetenschappelijke geschiedenis, met de mensen van alle continenten. Nou denkt u misschien, hoe dan? Hoe kan het onderwijs verbeterd worden? Laat me een voorbeeld geven. Wie ontdekte Amerika?

DE ONTDEKKER VAN AMERIKA

Christopher Columbus. Dat leerde ik ook bij geschiedenis. Onze dertienjarige dochter, Riana, leerde het dit jaar nog. Maar, als je dezelfde vraag stelt aan de Siksika, zullen ze dit antwoord niet geven. Hun voorouders waren al lang voordat Columbus arriveerde op het continent aanwezig. Columbus wist dat zelf ook wel: Hij werd zelfs welkom werd geheten toen hij aan land kwam. De geschiedenis van Amerika begint niet bij Columbus: Dat is incorrect en beledigend voor de oorspronkelijke bewoners.[3] Een gedeelde geschiedenis, dus.

RECHTE DRIEHOEKEN

Een ander voorbeeld, nu uit de wiskunde: Laten we een rechthoekige driehoek nemen met twee rechte zijden van lengte *a* en *b* en een schuine zijde van lengte *c*. We leren op school dat de lengte van de zijden de volgende relatie hebben: $a^2 + b^2 = c^2$. We leren dat dit de stelling van Pythagoras is. Alleen, deze stelling was al veel langer bekend. De oude Egyptenaren kenden de stelling al, en de stelling was zelfs nog eerder bekend in Mesopotamië.[4] De stelling van Pythagoras was dus alleen voor de Grieken nieuw. Een gedeelde geschiedenis dus.

DE PIRAMIDE VAN MASLOW

Nog een laatste voorbeeld, deze keer uit de psychologie: Dit is de piramide van Maslow. Abraham Maslow stelt dat er een hiërarchie van behoeften is. Als je behoeften hoger in de hiërarchie wilt vervullen (zoals eigenwaarde en respect voor anderen) dan is het noodzakelijk om eerst de behoeften lager in de hiërarchie te vervullen (zoals lichamelijke veiligheid en werkzekerheid). Als je *dit* laat zien aan de Siksika, zullen ze vertellen dat zo'n hiërarchie van behoeften een belangrijk plaats inneemt binnen hun cultuur. Dat is toevallig! Of misschien toch niet. In de zomer van 1938 bracht Maslow langere tijd door bij de Siksika. Je zou kunnen zeggen dat Maslow stage heeft gelopen bij de Siksika.[5] Wederom een gedeelde geschiedenis.

WETENSCHAPSGESCHIEDENIS EN HET ONDERWIJS

Ik begrijp wel dat het niet mijn taak is om de wetenschapsgeschiedenis te herschrijven. Maar we kunnen in elk geval studenten het besef bijbrengen dat er een gedeelde geschiedenis bestaat, waarvan een deel ontbreekt. Mekka Okereke, software engineer bij Google, stelt daarom voor om aan elk voorbeeld uit onze "westerse" wetenschapsgeschiedenis de volgende drie woorden toe te voegen: "*voor witte mensen.*"[6] Dus, Columbus ontdekte Amerika *voor witte mensen*. Pythagoras ontdekte de stelling voor rechte driehoeken *voor witte mensen*. Maslow ontdekte de hiërarchie van behoeften, u begrijpt het al, *voor witte mensen*. En misschien moeten we daar "voor witte mannen" van maken, want ook de wetenschappelijke bijdragen van vrouwen worden in de wetenschaps-geschiedenis stelselmatig onderbelicht.[7] Maar, dat is een onderwerp voor een andere rede. Geachte rector, ik ben nu bijna zover dat ik mijn rede over zoekmachines kan beginnen, maar staat u mij toe om mijn rede in te leiden met nog een klein stukje wetenschapsgeschiedenis?

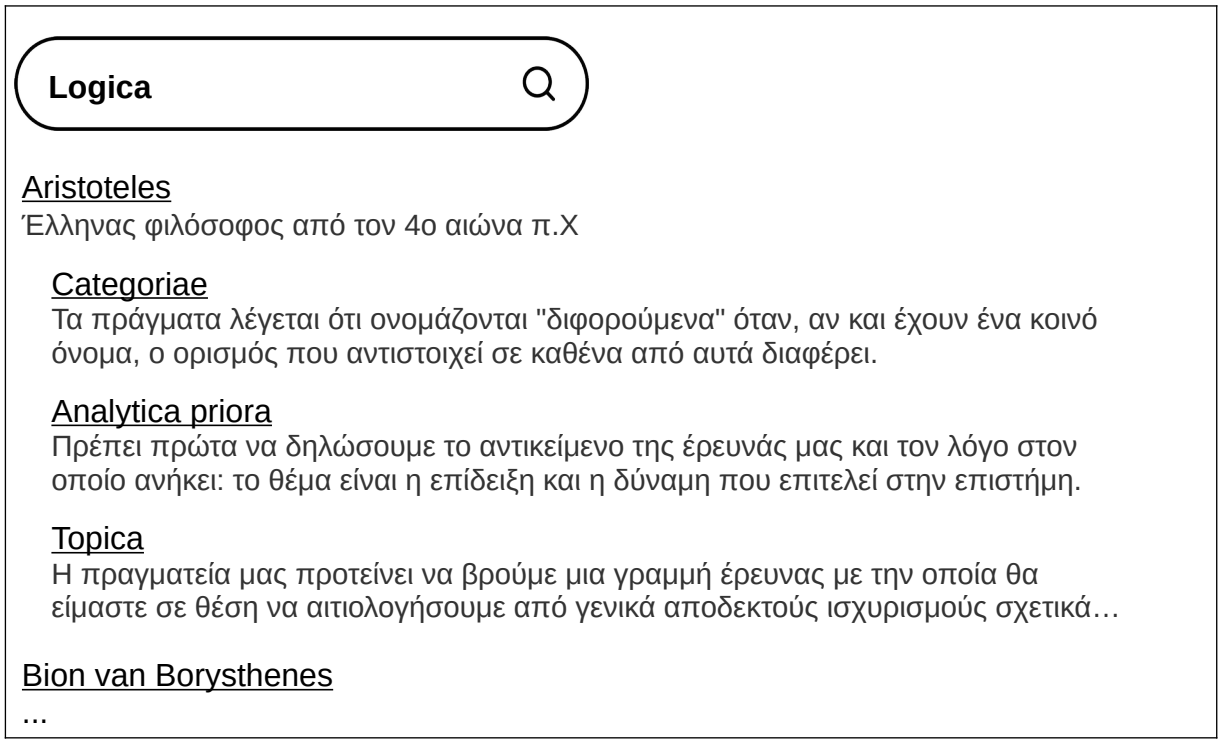
DE ONTDEKKING VAN METADATA

We gaan hiervoor naar de zevende eeuw voor het begin van onze jaartelling, naar de stad Nineveh. Nineveh lag waar nu de stad Mosul ligt in Irak. Nineveh stond in die tijd bekend als een uitstekende plek om te studeren, net als nu Nijmegen, zou je kunnen zeggen. Dat kwam door Nineveh's enorme bibliotheek. Er wordt geschat dat de bibliotheek meer dan 100.000 kleitabletten bevatte.[8] Stel je zoekt informatie in de bibliotheek van Nineveh. Hoe vind je het juiste kleitablet in zalen en zalen vol met 100.000 kleitabletten? Om daarbij te helpen, werden kleitabletten gesorteerd naar categorie: wiskunde in zaal 1, geschiedenis in zaal 2, filosofie in zaal 3, enz. Zo'n categorie is eigenlijk extra informatie die aan een kleitablet wordt toegevoegd. We noemen dat ook wel *metadata*, data over data.

DE PINAKES

Ongeveer vier eeuwen later was de technologie om teksten efficiënt op te slaan verbeterd door het gebruik van papyrus. De bibliotheek van Alexandrië, in het huidige Egypte, bevatte naar schatting maar liefst 500.000 papyrusrollen. We weten nog wie de eerste bibliothecaris was van de biblio-theek: een Griekse geleerde genaamd Zenodotus. Zenodotus was de eerste persoon waarvan we weten dat hij categorieën toekende aan papyrusrollen, zeg maar de ontdekker van metadata (voor

witte mensen). Om informatie sneller te kunnen vinden, maakte de Griekse dichter Callimachus de Pinakes, een lange lijst waarin hij de papyrusrollen van de bibliotheek van Alexandrië uitgebreid beschreef, een uitgebreide lijst met metadata, dus.[9] De Pinakes zelf zijn verloren gegaan, maar we weten uit beschrijvingen dat de Pinakes er ongeveer uit hebben gezien als in figuur 1.[10]

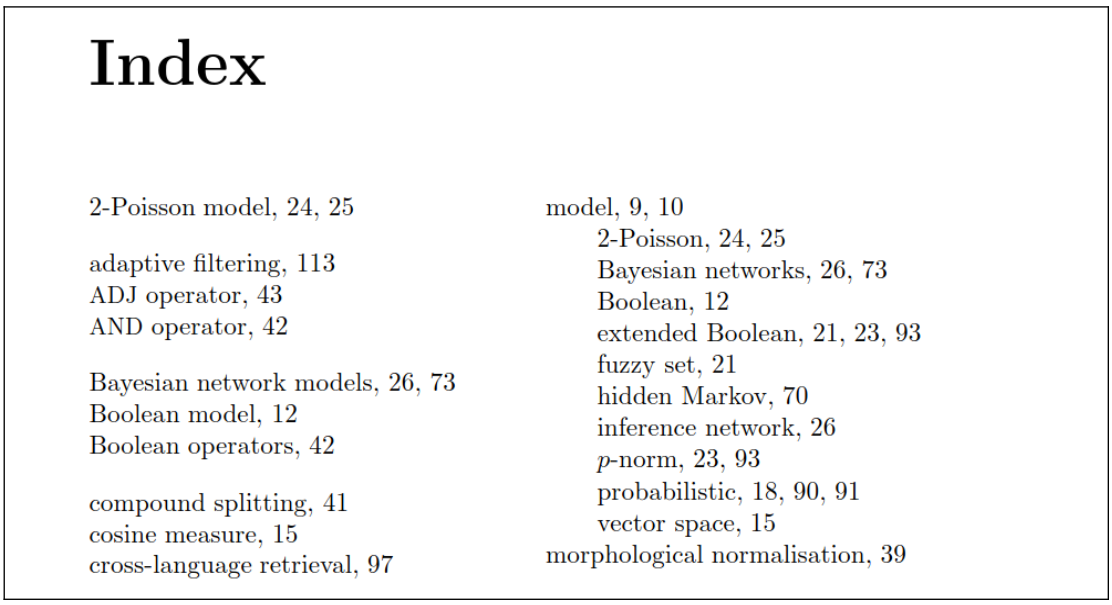


Figuur 1, impressie van de Pinakes

Voor elke categorie, laten we zeggen, “Logica” voegde Callimachus de auteurs toe op alfabetische volgorde, te beginnen met de ‘A’ van, bijvoorbeeld, Aristoteles. Voor elke auteur voegde hij een kleine biografie toe. Vervolgens voegde hij de titel en de openingsregel van elk werk toe dat de auteur over logica had geschreven. Aristoteles schreef bijvoorbeeld *Categoriae*, *Analytica priora* en *Topica*. Daarna kwam de volgende auteur, bijvoorbeeld met de ‘B’ van Bion van Borysthenes. Met een beetje fantasie zie je hier een pagina met zoekresultaten van een zoekmachine in.[11] De Pinakes zelf bestonden uit 120 delen, dus je kunt je voorstellen dat het nog steeds even kon duren voordat je de juiste papyrusrol had gevonden.

DE INDEX

Om nog sneller informatie te kunnen vinden in een catalogus zoals de Pinakes, voegen we tegenwoordig vaak een index toe. Figuur 2 bevat het begin van de index bij mijn proefschrift.[12] De index bestaat uit twee delen: Ten eerste een lijst met alle categorieën, ook wel lemma’s genoemd, en ten tweede voor elk lemma een lijst met paginanummers waar ik informatie over het lemma kan lezen. Nu kunnen we dus het lemma “Logica” opzoeken in de index en weten we meteen in welke van de 120 delen van de Pinakes we moeten wezen, of in mijn proefschrift, als ik meer wil weten over het “2-Poisson model” dan moet ik naar pagina’s 24 en 25.



Index

Figuur 2, voorbeeld van een index

DE BESTE INDEX

Maar hoe bepaal je nu welke lemma's je moet gebruiken in de index, en hoe bepaal je welke pagina's belangrijk zijn? Een van de eerste wetenschappers die deze vraag probeerde te beantwoorden was Cyril Cleverdon van de Universiteit van Cranfield in Engeland in de jaren '60 van de vorige eeuw. Cleverdon vroeg aan proefpersonen om de zoekresultaten van zoekmachines met verschillende indexen te beoordelen die allemaal de beste werkwijzen uit de bibliotheekwetenschap gebruikten. Cleverdon voegde nog één index daaraan toe, een index waarin elk woord dat in een tekst voorkomt aan de index wordt toegevoegd, met bij ieder woord elke pagina waarop dat woord voorkomt. Wat bleek? De proefpersonen waren het meest tevreden met de resultaten van die laatste index (elk woord op elke pagina). Cleverdon was daar zo verbaasd over dat hij schreef:

> *Deze conclusie is zo controversieel en zo onverwacht (…) ze lijkt in strijd te zijn met alles waarmee we als bibliothecarissen zijn opgeleid.*[13]

Logisch dat Cleverdon geschokt was, want: Voor een bibliothecaris is een index met elk woord en elke pagina natuurlijk niet werkbaar, maar voor computers wel. Computers kunnen heel snel, heel veel simpele instructies uitvoeren. Voor de computer is de index met elk woord op elke pagina de beste index. We hebben geen bibliothecaris meer nodig, ook niet om de index te maken.

SAMENVATTING

We hebben nu alle ingrediënten voor een zoekmachine: Het wereldwijde web is de grootste bibliotheek in de geschiedenis met miljarden en miljarden webpagina's. Als we deze pagina's downloaden kunnen we de *metadata*, zoals de titel en openingsregels, gemakkelijk in elke pagina vinden en in een lange lijst zetten. De *index* maken we daarna door elk woord als lemma op te nemen, en voor elke lemma alle pagina's op te sommen waarin het woord voorkomt. De zoekmachine kan nu razendsnel de juiste pagina's vinden voor elke zoekvraag.
Mijnheer de rector, ik ben nu klaar om mijn rede over zoekmachines uit te spreken.

ONBEPERKTE TOEGANG TOT ALLE INFORMATIE DOOR SAMEN TE WERKEN

Het is mijn droom om onbeperkte toegang tot alle online informatie mogelijk te maken, met behulp van zoekmachines die *samen* worden ontwikkeld en onderhouden. Laat ik context geven bij mijn droom. Er zijn op dit moment slechts vier grote zoekmachines met een eigen webindex: Google en

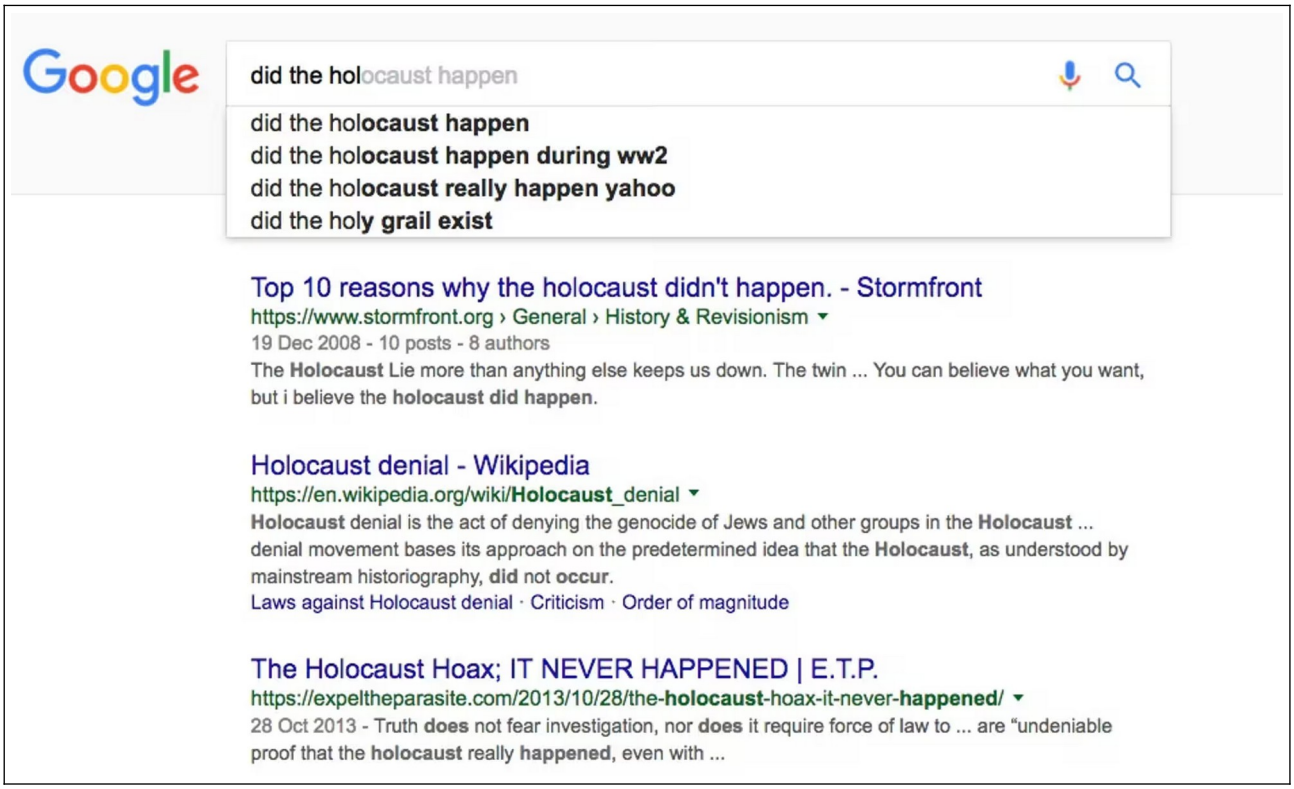
Microsoft Bing (beiden in de Verenigde Staten van Amerika), Baidu (in China) en Yandex (in Rusland). Deze zoekmachines beweren ook dat ze “onbeperkte toegang tot alle informatie” als doel hebben, maar niet door samen te werken. Integendeel, de bedrijven achter deze zoekmachines hebben een dubbele agenda. Naast het toegankelijk maken van informatie, willen ze namelijk ook zoveel mogelijk privégegevens verzamelen. Dit doen ze om mensen gerichte advertenties te laten zien, of om mensen op een andere manier te manipuleren. In plaats van samen te werken, streven ze ernaar om alle aspecten van het zoeken te controleren.

ZOEKMACHINES EN MEER

Wat zijn dan die aspecten? Zoekmachines halen webpagina’s op, ze halen metadata uit de pagina’s, ze bouwen de index, ze zoeken in de index, en ze laten de zoekresultaten zien. Maar daarnaast doen ze ook allerlei dingen die niks met zoekmachines te maken hebben: Ze bieden e-mail aan; video’s, nieuws, plattegronden en routeplanning, chat, web analytics, enz., zodat ze precies kunnen bijhouden wat mensen iedereen doet. Ze leveren zelfs de technologie die nodig is om hun zoekmachines te gebruiken: namelijk de webbrowsers (zoals Chrome en Edge) en besturings-systemen (zoals Windows en Android). Alles om zo gericht mogelijk advertenties te kunnen verkopen.

ADVERTENTIES EN ZOEKMACHINES

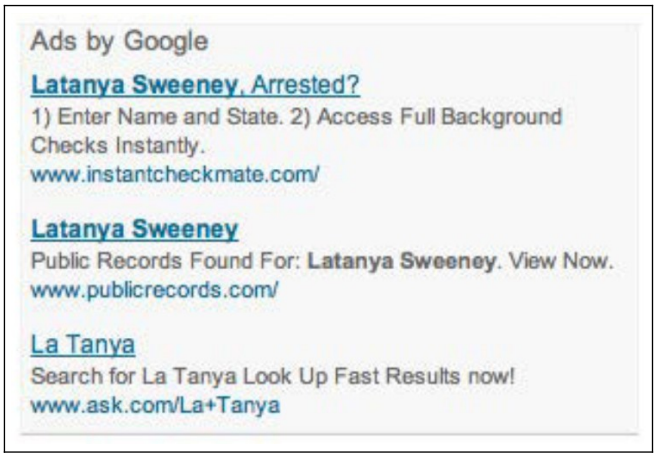
Waarom is dat erg? Stel je bent een zoekmachine *en* je verkoopt ook advertenties. Iemand zoekt naar informatie over de holocaust, en je kunt kiezen om betrouwbare informatie te laten zien van een website zonder advertenties *of* onbetrouwbare informatie van een website die bij jou advertenties koopt. Betrouwbare informatie tonen of geld verdienen? Wat zou je doen?



Figuur 3, een voorbeeld van immorele zoekresultaten[14]

Figuur 3 laat een schermafdruk zien uit een onderzoek van onderzoeksjournalist Carole Cadwal-ladr.[14] Alhoewel de gebruiker misschien naar iets heel anders op zoek was, suggereert de

zoekmachine “Did the holocaust happen?” (heeft de holocaust plaatsgevonden?)[15] Misschien wel omdat iemand daarvoor advertentiegeld betalen wil. Als je vervolgens deze zoekvraag kiest dan is het eerste resultaat de neonazi website Stormfront; misschien wel omdat die site ook advertenties gebruikt. Voordat je het weet, zet je als zoekmachine aan tot genocide. Dit is niet een denkbeeldig probleem. Dat weten we ook van Facebook’s rol bij het aanzetten tot geweld tegen de Rohingya in Myanmar.[16] *Daarom* is het erg dat zoekmachinebedrijven ook advertentiebedrijven zijn.



Figuur 4, een voorbeeld van immorele advertenties[17]

Nog een voorbeeld. Stel je bent eigenlijk een advertentiebedrijf. Je wilt dat mensen zoveel mogelijk op advertenties klikken, want aan elke klik verdien je geld. Je gebruikt daarom een zelflerend systeem dat de advertenties laat zien waar het meest waarschijnlijk op geklikt gaat worden. Na een tijdje krijg je klachten van mensen die bij hun naam suggestieve advertenties krijgen, zoals die in figuur 4 voor Latanya Sweeney: “Latanya Sweeney arrested”? Werd Latanya Sweeney gearresteerd? Nee. Het blijkt bovendien dat de meeste klachten komen van mensen met een donkere huidskleur. Of beter gezegd, mensen die een naam hebben die vaak gebruikt wordt in families die afstammen van tot slaaf gemaakten uit Afrika, namen zoals Latanya, of Darnell, of Leroy.

Latanya Sweeney bestaat echt. Ze is professor aan de universiteit van Harvard in de Verenigde Staten en ze onderzocht de Google resultaten voor meer dan 2000 namen die ofwel meer in witte families of meer in zwarte families worden gebruikt.[17] Wat blijkt? Google toonde systematisch discriminerende advertenties bij “zwarte” namen. De verklaring hiervoor is dat de meeste gebruikers van de zoekmachines zich racistisch gedragen. Ze klikken bij de naam Latanya meer op *arrestatie* advertenties. Google is niet bewust racistisch, maar winstmaximalisatie levert onbedoeld een discriminerende, racistische zoekmachine. *Daarom* is het erg dat zoekmachinebedrijven ook advertentiebedrijven mogen zijn.

ONDETECTEERBARE MANIPULATIE

En het is waarschijnlijk nog erger dan we denken. De meeste oneerlijke praktijken van zoek-machines blijven waarschijnlijk “onder de radar”, onopgemerkt. Collega Tim de Jonge ontwikkelde vorig jaar UNFair, de “Undetectably Nefarious Fair Ranker”, ofwel een zoekmachine die eerlijk lijkt, maar ondetecteerbaar toch schandalige, manipulatieve resultaten laat zien.[18] Dat kunnen zoekmachines doen door gepersonaliseerde resultaten te geven, waarbij alleen mensen die makkelijk te beïnvloeden zijn, bepaalde resultaten te zien krijgen. Tims conclusie? Als zoek-machines willen, kunnen ze ons manipuleren zonder dat onderzoekers zoals Cadwalladr of Sweeney het door hebben. Het is daarom nog belangrijker om elke stimulans om ons te manipuleren bij zoekmachines weg te nemen. Wat is er nodig om dit op te lossen?

WAT TE DOEN?

Ten eerste *wet- en regelgeving*. Een onderneming zou niet een zoekmachine mogen aanbieden, en ook nog advertenties. Wetgeving is niet mijn specialiteit. Daarom is het fantastisch dat de Radboud Universiteit een speciale onderzoekslab voor digitalisering en de samenleving heeft, genaamd iHub. In dit lab werken informatici, zoals ik, samen met rechtsgeleerden, zoals Frederik Zuiderveen Borgesius, die wel digitale wet- en regelgeving als specialiteit hebben. Wat er ten tweede nodig is, is *samenwerking*. Daar wil ik het nu over hebben. Wanneer wij voor zoekmachines kunnen laten zien dat meer samenwerking mogelijk is, dan wordt het voor de wetgever makkelijker om strengere regelgeving vast te stellen.

VIER MANIEREN OM SAMEN TE WERKEN

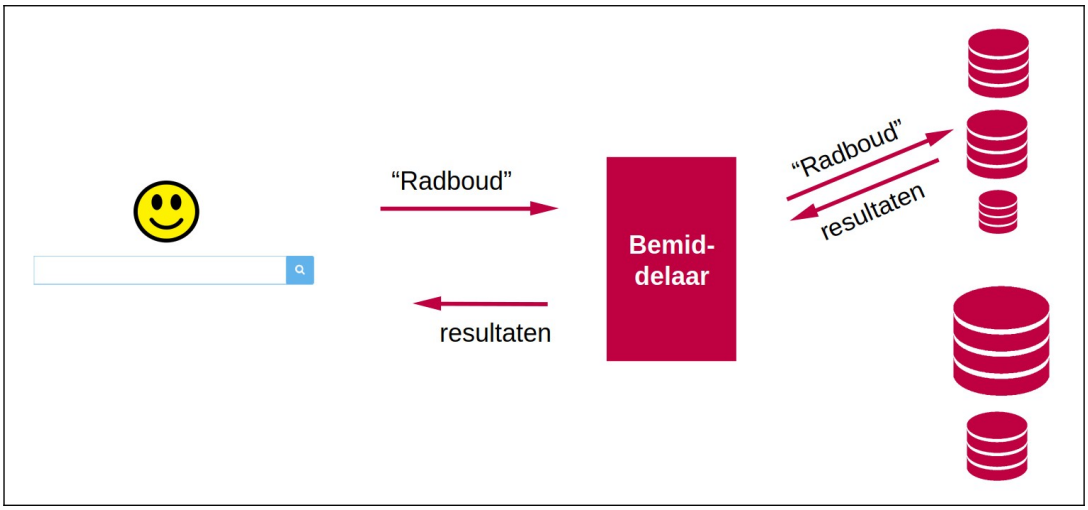
Bij ons onderzoek naar zoekmachines proberen we uit te vinden hoe we meer samenwerking mogelijk kunnen maken. Dat doen we op verschillende manieren. Ik zal vier manieren noemen waarop ik een verschil wil maken.

1. Laat bestaande zoekmachines samenwerken
2. Bouw gezamenlijke een zoekmachine
3. Nieuwe standaarden die samenwerken mogelijk maken, en
4. Samenwerken door gefaseerd zoeken

Laten we beginnen met punt 1. Stel, we maken helemaal niet één grote zoekmachine, maar in plaats daarvan zorgen we ervoor dat kleine bestaande zoekmachines kunnen samenwerken om toch het hele web te doorzoeken. Kan dat?

1. LAAT ZOEKMACHINES SAMENWERKEN

Deze vraag heb ik de afgelopen jaren uitgebreid mogen onderzoeken. Het idee wordt gevisualiseerd in figuur 5.



Figuur 5, samenwerkende zoekmachines

Rechts zien we meerdere zoekmachines, afgebeeld als tonnetjes.[19] Elke zoekmachine heeft een eigen lijst met metadata en een eigen index. Links zien we een webpagina met een zoekveld. Daartussen plaatsen we een bemiddelaar, die de zoekvraag ontvangt van de web pagina. Deze bemiddelaar stuurt de zoekvraag vervolgens door naar de juiste zoekmachines, rechts. De bemiddelaar moet hierbij een klein aantal zoekmachines selecteren om deze aanpak efficiënt te houden. Dit roept allerlei interessante onderzoeksproblemen op, waar we de afgelopen jaren oplossingen voor hebben gevonden.

Bijvoorbeeld,

- Dolf Trieschnigg[20] ontwikkelde de “search result finder” die voor elke willekeurige webpagina, bijvoorbeeld de pagina van de Radboud Universiteit, de zoekresultaten herkent;
- Sander Bockting[21], Sergio Duarte[22], Robin Aly[23] en Dong Nguyen[24] onderzochten hoe de bemiddelaar het beste kan bepalen wat de beste zoekmachines zijn;
- Almer Tigelaar[25] liet zien dat de bemiddelaar kan leren van een steekproef van de zoekresultaten en dat het onnodig is om alle metadata en indexen te delen;
- Kien Tjin-Kam-Jet[26] liet zien hoe de web pagina de resultaten van meerdere zoekmachines kan samenvoegen;
- Ke Zhou[27] liet zien hoe je de relevantie van zoekmachines in dit scenario kan afstemmen op de intentie van de gebruiker;
- Thomas Demeester[28] onderzocht welke resultaten mensen daadwerkelijk willen zien in dit scenario, door systematisch duizenden resultaten van meer dan 150 zoekmachines te laten beoordelen door studenten van de Universiteit Gent;
- Vorig jaar nog, liet Negin Ghasemi[29] zien hoe een kleine zoekmachine kan leren van een grote zoekmachine, bijvoorbeeld, hoe een nieuwe webwinkel in een klein land, automatisch kan leren van een bestaande webwinkel in een groot land.

SEARSIA

Ik ben van mening dat het onderling samenwerken van zoekmachines haalbaar is. Het is mogelijk om kleine zoekmachines te laten samenwerken om samen een grote zoekmachine te maken. Samen met Dolf Trieschnigg ontwikkelde ik Searsia, een speciale zoekmachine met bemiddelaar die veel van de zojuist genoemde oplossingen gebruikt.[30] Hiermee kan iedereen samenwerkende zoekmachines maken. Maar iedereen doet dit nog niet. Misschien is het niet de beste manier om samenwerking mogelijk te maken.

2. BOUW GEZAMENLIJK EEN ZOEKMACHINE

Daarom wil ik graag mijn tweede manier om samen te werken bespreken: Is het mogelijk om gezamenlijk een zoekmachine te bouwen? Bij veel complexe technologie is het zo dat fabrikanten met tientallen leveranciers samenwerken om een eindproduct te maken. Laten we als voorbeeld een auto nemen, bijvoorbeeld de Toyota Aygo. Het is niet zo dat Toyota alle onderdelen van de Aygo zelf maakt. De motor en chassis (ook wel de *power train* genoemd) wordt door Toyota gemaakt, maar andere fabrikanten hebben zich gespecialiseerd in andere onderdelen. Elke fabriek heeft zijn eigen toeleveranciers. FTS bouwt bijvoorbeeld benzinetanks voor Toyota; Sango bouwt de uitlaat en geluidsdempers; Sumitomo levert de elektrische bedrading, antennes en sensors; Taiho Kogyo levert de lagers en pakkingen; en zo kunnen we nog even doorgaan voor de banden, de bekleding, het dashboard, enzovoort.[31] De samenwerking in de auto industrie gaat zelfs zover dat verschillende fabrikanten samen auto’s ontwerpen. Toyota ontwierp de Aygo samen met Peugeot en Citroën. De Aygo heeft hetzelfde basisontwerp als de Peugeot 108 en de Citroën C1. Dat is nog eens samenwerken!

Zoals we net al gezien hebben, zijn ook zoekmachines complex. Voordat we een zoekmachine hebben, moeten we eerst alle webpagina’s ophalen (de zogenaamde crawler), dan moeten we de pagina’s analyseren en metadata opslaan; daarna moeten de we index maken; vervolgens is er een onderdeel dat de pagina’s daadwerkelijk zoekt, een onderdeel dat de suggesties levert, een onderdeel dat spelfouten opspoort en verwijdert, enzovoort, enzovoort. Grote zoekmachines zoals Google, Bing, of Baidu maken al deze onderdelen zelf. Is het logisch dat de grote zoekmachines bijna alles zelf doen? Misschien niet. Zou het mogelijk zijn om complexe software, en complexe tekstanalyse te organiseren zoals de auto industrie (of meer in het algemeen, zoals de assemblage-

industrie)? Samen met andere universiteiten en onderzoeksinstituten in Europa zoeken we aan de Radboud Universiteit naar antwoorden op deze vragen.

OPEN WEB SEARCH

We doen dit in het Europese onderzoeksproject Open Web Search. Het doel van het project is om de markt voor zoekmachines open te breken, door te laten zien dat het mogelijk is om samen een zoekmachine te maken. Verschillende partners leveren verschillende stukjes van de zoekmachine-puzzel. Dat gaat in het project als volgt: De Universiteit van Passau in Duitsland ontwikkelt een manier om gezamenlijk het web te crawlen, dat wil zeggen, gezamenlijk pagina's te ontdekken en op te halen. De Universiteit van Leipzig, ook in Duitsland, ontwikkelt gereedschappen om de pagina's te analyseren en verwerken tot metadata. Vervolgens maken wij in Nijmegen de index. Tenslotte, gebruikt de Universiteit van Graz, in Oostenrijk, de index en de metadata om het zoeken mogelijk te maken. Deze halffabricaten worden als puzzelstukjes "in elkaar gezet" bij grote Europese rekencentra, zoals die van het Zwitserse CERN, het Tsjechische IT4I en het Finse CSC. Het doel is om niet alleen de software te delen om de halffabricaten te maken, maar ook de halffabricaten zelf: De webpagina's, de metadata, en met name de index: Een open web index.[32]

THIRD-PARTY CALLS

En iedereen kan meedoen. Om te laten zien dat het mogelijk is om samen zoekmachines te maken, nodigen we bedrijven en instellingen uit om onze infrastructuur, onderdelen en halffabricaten te gebruiken om zoekmachines te maken zoals we auto's maken. We noemen dit de "third-party calls". Ben je zo'n derde partij en heb je een goed voorstel, dan krijg je van ons budget om dat voorstel te realiseren. En je hoeft echt niet meteen een alternatief voor Google te maken. Het kan bijvoorbeeld een zoekmachine voor minder validen zijn, of een zoekmachine voor kinderen.[33]

3. STANDARDISATIE

Laten we kijken naar de derde manier van samenwerken: Standaardisatie. Stel, je wil je nieuwe iPhone 19 Pro Max Plus SE opladen. Apple is berucht om zijn pogingen om samenwerking te frustreren, bijvoorbeeld door ervoor te zorgen dat iPhones moeilijk te repareren zijn, en door elk nieuwe model een andere oplader en andere koptelefoons te geven. Dus, ja, welke oplader moet ik nu voor deze iPhone 19 gebruiken? Gelukkig heeft de EU enkele jaren geleden besloten dat per 2024 alle draagbare apparaten een USB-C oplader moeten kunnen gebruiken.[34] Standaarden, zeker verplichte standaarden, zoals deze van de EU zijn enorm belangrijk om monopolies open te breken en samenwerking mogelijk te maken. Zou het niet mooi zijn als de EU over enkele jaren het volgende bepaalt: Wanneer je een zoekmachine maakt, dan *moet* je de index beschikbaar maken in een door de EU goedgekeurd standaard format. Een dergelijke standaard ontwikkelen we op dit moment aan de Radboud Universiteit.

Mijn collega's, Arjen de Vries en Chris Kamphuis, stelden in 2020 de CIFF standaard voor. Dat deden ze met de ontwikkelaars van de software van in totaal vier zoekmachines.[35] CIFF is een afkorting voor "Common Index File Format" ofwel het "gemeenschappelijke indexbestands-formaat". CIFF was oorspronkelijk bedoeld om resultaten voor onderzoek te delen, maar nu is ons doel groter. De standaard moet een breedgedragen standaard worden, die ook werkt bij onverwachte combinaties van onderdelen en halffabricaten. Zo heeft Gijs Hendriksen vorig jaar een uitbreiding van de standaard getest, waardoor het makkelijker moet worden om indexen te delen die ongebruikelijke woorden, of lemma's, gebruikt.[36] Dit is belangrijk voor het verwerken van teksten in talen die niet ons alfabet gebruiken, zoals Chinees en Japans. Het maakt de CIFF standaard in de toekomst hopelijk een universele standaard voor het delen van indexen, zoals de USB-C stekker is voor opladers.

4. HET ZOEKEN IN FASES

Dan zijn we aangekomen bij mijn vierde en laatste manier om samen te werken. Het zoeken in meerdere fases. Zoeken in meerdere fases wordt al zeker 20 jaar toegepast.[37] De eerste fase werkt zoals we gezien hebben met een index, maar in plaats van 10 resultaten op te halen (genoeg voor 1 pagina met resultaten), worden er nu veel meer resultaten opgehaald, misschien wel 1000. De tweede fase gebruikt geen index, maar een complex sorteeralgoritme dat automatisch leert van voorbeelden. Dit sorteeralgoritme rangschikt de 1000 resultaten van fase 1 opnieuw. Dit levert 10 andere, hopelijk betere, resultaten op.

GROTE TAALMODELLEN

Samenwerken is relatief makkelijk bij gefaseerd zoeken. Één leverancier levert de metadata en index voor fase 1 (de “first-stage retriever”) en een andere leverancier levert het sorteeralgoritme (de “reranker”) voor fase 2. Voor fase 2 worden sinds een jaar of vijf vaak zogenaamde “large language models” gebruikt, *grote taalmodellen* dus. Dit zijn modellen die kunnen ‘leren’ van enorme hoeveelheden tekst. Het leren van deze grote taalmodellen kost enorm veel tijd en geld, dus ze worden vooral gemaakt door grote bedrijven als Google, Facebook en OpenAI. Rodrigo Nogueira en collega’s van New York University gebruikten in 2019 BERT, een groot taalmodel van Google.[38] Ze lieten zien dat BERT de kwaliteit van zoekresultaten kon verbeteren van 19% voor fase 1 naar 36% na fase 2, bijna tweemaal zoveel goede resultaten dus.[39] Dat is indrukwekkend. Heel veel onderzoekers, ook wij hier aan de Radboud Universiteit, zijn toen ook aan de slag gegaan met zulke grote taalmodellen. Sterker nog, de hele wereld is hier mee bezig. GPT is bijvoorbeeld een groot taalmodel dat de basis vormt voor ChatGPT, daar heeft u vast wel iets over gehoord in de media.[40]

HET ENERGIEGEBRUIK VAN GROTE TAALMODELLEN

Om deze modellen te kunnen gebruiken moeten we speciale supercomputers, zogenaamde GPU clusters gebruiken. De hoeveelheid energie die dit kost is enorm, zoveel dat ze gekoeld moeten worden met water. Het blijkt dat fase 2 in het experiment van Noguiera meer dan 138.000 keer zoveel energie kost dan fase 1.[41] Gefaseerd zoeken met grote taalmodellen kost absurd veel energie. Het is alsof ik een manier voorstel om beter van Nijmegen naar Amsterdam te reizen. 19% van de mensen reist graag met de auto en 36% reist graag met de *Space Shuttle*! Dat snap ik wel. Dat lijkt me ook leuk.

DE SPACE SHUTTLE

Dit lijkt een absurde vergelijking, maar dat is het niet. De Space Shuttle heeft een enorme externe tank, waarin ongeveer anderhalf miljoen liter raketvloeistof past. Dat is ongeveer 100.000 keer zoveel als nodig is voor een ritje Nijmegen Amsterdam met de auto. En denk maar niet dat u met de Space Shuttle *sneller* in Amsterdam bent. Je gaat namelijk eerst recht omhoog. Ook de gefaseerde zoekmachine is niet sneller. Fase 2 kost namelijk alleen maar extra tijd. Daarnaast levert fase 2 vooral meer uitstoot. Als we uitrekenen hoeveel CO_2 wordt uitgestoten in fase 1 en fase 2 van het experiment, dan is dat 50 miligram CO_2 voor fase 1 en meer dan 8 *kilo*gram CO_2 voor fase 2 (als we stroom van het Nederlandse net gebruiken). Dit is totaal onverantwoord.

VERANTWOORD GEBRUIK VAN AI

We leven in een tijd waarin klimaatverandering merkbaar is geworden. Voor sommige landen betekent klimaatverandering dat steeds meer mensen zullen sterven tijdens hittegolven. Voor andere

landen betekent dit dat steeds meer mensen zullen sterven door mislukte oogsten. Nederland zal er op een gegeven moment simpelweg niet meer zijn door de stijgende zeespiegel. We *moeten* nu stoppen met fossiele brandstoffen en dat kan alleen als we *nu* ons energiegebruik zoveel mogelijk beperken. We moeten daarom stoppen met het domweg gebruiken van grote taalmodellen in ons onderzoek, en ja, u zou daarom ook moeten stoppen met het gebruiken van chatGPT.

DUURZAAM VOORUIT

Voor mijn promotieonderzoek – ik promoveerde in 2001 bij Franciska de Jong – werkte ik ook aan taalmodellen voor zoekmachines.[12] Deze taalmodellen moeten we nu maar *kleine* taalmodellen noemen, om verwarring te voorkomen met de grote taalmodellen zoals BERT en GPT. Samen met Wessel Kraaij, Arjen de Vries en Thijs Westerveld konden we de kleine taalmodellen gebruiken in bestaande zoekmachines, gewoon in fase 1. De algoritmes van kleine taalmodellen kun je namelijk wiskundig zo herschrijven dat ze direct uit de index berekend worden. Ze kosten daarom geen extra energie. De grote uitdaging voor het huidige zoekmachineonderzoek is om met oplossingen te komen waarmee we de kwaliteitsverbeteringen van grote taalmodellen ook kunnen gebruiken in fase 1, zonder dat het absurd veel energie kost.

CONCLUSIE

Mijn conclusies vandaag zijn de volgende. Mijn hoop is gevestigd op samenwerken op een wijze manier, de manier van de Siksika, *waarbij het resultaat verbeterd door hulpmiddelen te delen, niet door steeds meer hulpmiddelen toe te voegen*.

1. Laten we zoekmachines maken zoals we auto’s maken, zodat er duizenden zoekmachines van de “Open Web Search lopende band” kunnen rollen.
2. Laten we zoekstandaarden ontwikkelen zodat de wetgever in de toekomst kan bepalen dat de index van de zoekmachine gedeeld *moet* worden als open web index. Zo’n open web index maakt niet alleen samenwerking mogelijk, maar ook meer onderzoek naar eerlijke zoekmachines die niet discrimineren.
3. Laten we heel kritisch zijn bij het gebruik van grote taalmodellen zoals BERT en GPT en oplossingen afwijzen wanneer ze te veel energie gebruiken, voor onze toekomst en de toekomst van onze kinderen en kleinkinderen
4. Tenslotte, laten we onze studenten leren over de rijke, gedeelde geschiedenis van de wetenschap, de geschiedenis van alle mensen van alle continenten. Het zoeken naar informatie is van alle tijden en alle culturen.

DANKWOORD

Voor ik mijn rede afsluit wil ik graag iedereen hier aanwezig bedanken. De Radboud Universiteit wil ik bedanken voor het vertrouwen in mij, in het bijzonder het Instituut voor Computing en Information Sciences, iCIS. Dank je wel, Tom Heskes, Elena Marchiori, David van Leeuwen en Martha Larson. Dank je wel Arjen de Vries, dat je me hierheen hebt gehaald, dat we naast directe collega’s ook vrienden zijn. Ik zou graag iedereen met wie ik heb samengewerkt persoonlijk willen bedanken, maar dat zou te lang duren. Onderzoek doen is namelijk ook, vooral *samenwerken*. Ik heb al de mensen genoemd die aan de samenwerkende zoekmachines hebben meegewerkt: Almer, Dolf, Dong, Ke, Kien, Robin, Sander, Sergio en Thomas. Jullie invloed was heel direct. Mijn leerstoel heet namelijk “Federated Search”! Mijn eerste promovendi, Vojkan Mihaljlovic en Henning Rode wil ik bedanken, misschien heb ik nooit gezegd hoe trots ik op jullie werk ben. Bedankt. Een enkele promovendus is zo succesvol dat de rollen inmiddels zijn omgedraaid. Dank je wel, Claudia Hauff, dat ik nu papers mag schrijven begeleid door jou.[42] Suzan Verberne wil ik bedanken voor samenwerking aan papers en onderwijs binnen SIKS. Claudia en Suzan: Eén van jullie had hier eigenlijk moeten staan, want we hebben meer vrouwelijke hoogleraren nodig. Ariana

Need, Michel Ehrenhard, Anna Priante, en Tijs van den Broek wil ik bedanken voor het mij wegslepen van monodisciplinair naar interdisciplinair onderzoek en echt teamwerk, dat had ik nodig. Franciska de Jong, Wessel Kraaij, Thijs Westerveld, Theo Huibers, en ook weer Arjen de Vries: Jullie stonden allemaal aan de wieg van mijn carrière, bedankt voor jullie steun. Bedankt ook Maarten de Rijke, je stond weliswaar niet aan die wieg, maar je steun was er op cruciale momenten. Mijn huidige promovendi: Negin Ghasemi, Tim de Jonge, Norman Knyazev, Benard Wanjiru, Gijs Hendriksen: Bedankt voor de prettige samenwerking. Ik leer veel van jullie. Dat begeleiden doe ik niet alleen, maar met heel veel fijne collega's: Arjen de Vries, Harrie Oosterhuis, Faegheh Hasibi, Frederik Zuiderveen Borgesius, en Patrick van Bommel.
Speciale dank gaat uit naar Marijn Hiemstra: Ik ben trots op wat je bereikt hebt, Marijn. Van jouw kennis van autotechniek heeft iedereen vandaag wat geleerd. Ik ook.
Tot slot wil ik ons gezin bedanken, Riana, Hidde en Jonah en vooral Barbara. Barbara, je zette de puntjes op de i voor mijn rede, je luisterde tot vervelens toe naar elke nieuwe versie, maar vooral: Je bent mijn grote liefde, steun en toeverlaat. Dank je wel.
Ik heb gezegd.

NOTEN

1 Het uitschakelen van concurrentie is een (onethische) strategie om marktdominantie te creëren en behouden, zie: Ioannis Lianos and Evgenia Motchenkova (2013). Market dominance and search quality in the search engine market. *Journal of Competition Law & Economics*, 9(2), 419-455.

2 Veroordeling Wilders wegens groepsbelediging blijft in stand. https://www.hogeraad.nl/actueel/nieuwsoverzicht/2021/juli/veroordeling-wilders-wegens-groepsbelediging-blijft-stand/

3 Bigelow, B. (1989). Discovering Columbus: Rereading the past. *Language Arts*, *66*(6), 635-643.

4 Maor, E. (2019). *The Pythagorean theorem: a 4,000-year history*. Princeton University Press.

5 Eaton, S. E. (2024). Decolonizing academic integrity: knowledge caretaking as ethical practice. *Assessment & Evaluation in Higher Education*, 1-16.

6 Okereke, M. (2023). When we say "discovered," we really mean discovered "for white people." *Hachyderm Mastodon* https://hachyderm.io/@mekkaokereke/109926764331109740

7 Des Jardins, J. (2010). *The Madame Curie complex: The hidden history of women in science*. City University Ney York Feminist Press.

8 Fincke, J. C. (2003). The Babylonian Texts of Nineveh: Report on the British Museum's Ashurbanipal Library Project. *Archiv für Orientforschung*, 111-149.

9 Sanderson, M. & Croft, W. B. (2012). The History of Information Retrieval Research. *Proceedings of the IEEE 100 (Special Centennial Issue)*, 1444-1451

10 Phillips, H. (2010). Great library of Alexandria. *Library philosophy and practice*.

11 Ik heb de fantasie een beetje geholpen door een zoekvak met een loepje om "Logica" te zetten.

12 Hiemstra, D. (2001). Using Language Models for Information Retrieval. *Proefschrift Universiteit Twente*.

13 Cleverdon, C., Mills, J., & Keen, M. (1966). Factors determining the performance of indexing systems. *The College of Aeronautics, Cranfield.*

14 Cadwalladr, C. (2016). Google is not 'just' a platform. It frames, shapes and distorts how we see the world. *The Guardian, 11 December 2016.*

15 De zoek vraag begint met “did the hol” en suggesties zijn mogelijke manieren om dat aan te vullen, zoals: “did the holy grail exist?”, “did the holidays start?”, “ did the holdovers win an oscar?”, “did the hole in the ozone close?”, enz.

16 Mozur, P. (2018). A Genocide Incited on Facebook, With Posts From Myanmar’s Military. *The New York Times*, *15 October 2018*.

17 Sweeney, L. (2013). Discrimination in online ad delivery. Communications of the ACM, 56(5), 44-54.

18 De Jonge, T. & Hiemstra, D. (2023). UNFair: Search Engine Manipulation, Undetectable by Amortized Inequity, *Proceedings of the International ACM Conference on Fairness, Accountability, and Transparency (FAccT)*.

19 De tonnetjes zijn eigenlijk harde schijven en worden vaak gebruikt om databases af te beelden.

20 Trieschnigg, D., Tjin-Kam-Jet, K., & Hiemstra, D. (2013) SearchResultFinder: Federated Search Made Easy. *Proceedings of the 36th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 113-1114

21 Bockting, S. & Hiemstra, D. (2009) Collection Selection with Highly Discriminative Keys. *Proceedings of the 7th Workshop on Large-Scale Distributed Systems for Information Retrieval (LSDS-IR), CEUR Workshop Series 480*. 9-16

22 Duarte Torres, S., Hiemstra, D. & Huibers, T. (2013). Vertical Selection in the Information Domain of Children, *Proceedings of the 13th Joint IEEE/ACM Conference on Digital Libraries (JCDL)*, 57-66

23 Aly, R., Hiemstra, D., & Demeester, T. (2013). Taily: Shard Selection Using the Tail of Score Distributions, *Proceedings of the 36th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 673-682

24 Nguyen, D., Demeester, T., Trieschnigg, D., & Hiemstra, D. (2016). Resource Selection for Federated Search on the Web. Technical Report TR-CTIT16-12, Centre for Telematics and Information Technology, University of Twente.

25 Tigelaar, A. & Hiemstra, D. (2010). Query-Based Sampling using Snippets, *Proceedings of the SIGIR Workshop on Large-Scale Distributed Systems for Information Retrieval, CEUR Workshop Proceedings 630*. 9-14

26 Tjin-Kam-Jet, K. & Hiemstra, D. (2010). Learning to Merge Search Results for Efficient Distributed Information Retrieval, *Proceedings of 10th Dutch-Belgian Information Retrieval Workshop (DIR)*

27 Zhou, K., Demeester , T., Nguyen, D., Hiemstra, D., & Trieschnigg, D. (2014) Aligning Vertical Collection Relevance with User Intent, In*Proceedings of the 23rd ACM Conference on Information and Knowledge Management (CIKM)*

28 Demeester, T., Trieschnigg, D., Zhou, K., Nguyen, D., & Hiemstra, D. (2015). FedWeb Greatest Hits: Presenting the New Test Collection for Federated Web Search”, *Proceedings of the 24th International World Wide Web Conference*

29 Ghasemi, N., Aliannejadi, M., Bonab, H., Kanoulas, E., de Vries, A., Allan, J., & Hiemstra, D. (2023). Cross-Market Product-Related Question Answering, *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*

30 Searsia. https://searsia.org

31 75 years of Toyota: List of member companies. https://www.toyota-global.com/company/history_of_toyota/75years/data/automotive_business/production/purchasing/member_companies/kyohokai.html

32 Granitzer, M., Voigt, S., Fathima, N.A., Golasowski, M., Guetl, C., Hecking, T., Hendriksen, G., Hiemstra, D., Martinovič, J., Mitrović, J., Mlakar, I., Moiras, S., Nussbaumer, A., Öster, P., Potthast, M., Srdič, M.S., Sharikadze, M., Slaninová, K., Stein, B., de Vries, A.P, Vondrák, V., Wagner, A., & Zerhoudi, S. (2024). Impact and development of an Open Web Index for open web search, *Journal of the American Society of Information Science and Technology 75(5)*, 512-520

33 Open Web Search launches new third-party calls: An invitation to researchers, innovators and computing centres to join the quest for a new Internet Search in Europe. https://openwebsearch.eu/third-party-calls-2024-media-release/

34 Yakimova, Y. (2022). Long-awaited common charger for mobile devices will be a reality in 2024. *European Parliament Press Release 4 October 2022*.

35 Lin, J., Mackenzie, J., Kamphuis, C., Macdonald, C., Mallia, A., Siedlaczek, M., Trotman, A., & de Vries, A.P. (2020). Supporting interoperability between open-source search engines with the common index file format. *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2149–2152

36 Hiemstra, D., Hendriksen, G., Kamphuis, C., & de Vries, A.P. (2023). Challenges of index exchange for search engine interoperability, *Proceedings of the fifth International Open Search Symposium (OSSYM)*

37 Liu, T.Y. (2009). Learning to rank for information retrieval. *Foundations and Trends in Information Retrieval*, *3*(3), 225-331

38 Devlin, J., Chang, M.W., Lee, K. & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 4171-4186

39 Nogueira, R., Yang, W., Cho, K., & Lin, J. (2019). Multi-stage document ranking with BERT. *arXiv preprint arXiv:1910.14424*.

40 Generatieve AI. Op deze pagina kun je meer lezen over Generatieve Artificiele Intelligentie zoals ChatGPT en de Radboud Universiteit. Deze pagina wordt regelmatig geactualiseerd. https://www.ru.nl/over-ons/nieuws-en-agenda/chatgpt

41 Scells, H., Zhuang, S., & Zuccon, G. (2022). Reduce, reuse, recycle: Green information retrieval research. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2825-2837

42 Hiemstra D., Hauff, C. & Azzopardi, L. (2017). Exploring the Query Halo Effect in Site Search: Leading People to Longer Queries, *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*