SAMENVATTING

Hoe kunnen zoekmachines de hyperlinks tussen documenten gebruiken om te bepalen welke documenten het meest relevant zijn voor een zoekvraag? Sommige zoekmachines gebruiken links om *populariteit* te meten, waarbij het onderliggende idee is dat een document (Webpagina) waar veel links naartoe wijzen populair is. Zoekresultaten worden geordend op basis van hun populariteit en op basis van de overeenkomst tussen de inhoud van een document en de zoekvraag. Een ander aspect van links is dat ze aangeven dat twee pagina's inhoudelijk iets met elkaar te maken hebben. Een hyperlink is tenslotte een verwijzing. Als pagina A relevant is voor een zoekvraag, dan zijn pagina's die gelinkt zijn aan A wellicht ook relevant. Linkinformatie bevat dus mogelijk bewijs voor de *inhoudelijke relevantie* van een document.

Dit proefschrift beschrijft onderzoek naar de waarde van die twee aspecten van linkinformatie—populariteit en inhoudelijke relevantie—voor het ordenen van zoekresultaten. Deze vraag werd eerder onderzocht binnen het vakgebied *information retrieval*. Vanaf eind jaren '90 werd op grote schaal geëxperimenteerd met het toepassen van linkinformatie om zoveel mogelijk relevante documenten te vinden voor een onderwerp en om deze documenten zo goed mogelijk te ordenen. De resultaten waren teleurstellend. Er waren geen consistente verbeteringen te meten door linkinformatie mee te nemen in het zoekproces. Vertegenwoordigers van zoekbedrijven gaven aan dat gebruikers van Webzoekmachines niet op zoek zijn naar zoveel mogelijk relevante informatie, maar naar de homepagina's van specifieke Websites, en pagina's die een goed startpunt vormen voor het verkennen van Webpagina's over een onderwerp. De onderzoekers besloten daarom hun aandacht te verschuiven naar deze Webspecifieke zoektaken, en hadden meteen succes. De homepagina's en andere belangrijke pagina's waar veel naar gezocht wordt zijn populaire pagina's, die met behulp van linkinformatie makkelijker te identificeren zijn. Hiermee leek de waarde van linkinformatie voor information retrieval duidelijk. Linkinformatie is nuttig voor het meten van populariteit, maar niet voor het meten van inhoudelijke relevantie.

De vraag waarom links niet nuttig zijn voor het meten van inhoudelijke relevantie werd nooit duidelijk beantwoord. Het doel van dit proefschrift is om de waarde van linkinformatie voor information re-

trieval preciezer en vollediger in kaart te brengen. Dit wordt eerst onderzocht met de Engelse *Wikipedia*, omdat deze in zijn geheel beschikbaar is, inclusief alle links tussen de Wikipediapagina's. Vanwege de encyclopedische aard is het zoeken naar informatie over een onderwerp in Wikipedia een natuurlijke taak. Dit maakt Wikipedia een geschikt startpunt voor het meten van linkbewijs voor inhoudelijke relevantie. De bevindingen worden vervolgens getoetst op een veel grotere collectie van Webpagina's, om vast te stellen in hoeverre zij generaliseerbaar zijn naar zoeken in het Web.

Bewijs voor populariteit kun je het best meten door alle links op het hele web te gebruiken en te tellen hoeveel er naar elke pagina gaan. Dit noemen we *globale* linkinformatie, afgeleid uit de *globale* linkstructuur. Voor populariteit is de richting van de link belangrijk; een link van A naar B maakt B populair, maar niet A. De pagina met de meeste links is het populairst. De volgorde waarin je de resultaten ordent wordt gedeeltelijk bepaald door de populariteit en gedeeltelijk door hoe goed de inhoud van een documenten overeen komt met de zoekvraag.

Bewijs voor inhoudelijke relevantie is af te leiden uit linkinformatie door eerst een lijst pagina's te zoeken waar de zoekwoorden vaak in voorkomen, en vervolgens alleen de links tussen die pagina's te gebruiken. Zo selecteer je links op een bepaald onderwerp: het onderwerp van je zoekvraag. Die lijst van pagina's noemen we de *lokale* pagina's en de links tussen die pagina's noemen we *lokale* links. Voor de inhoudsrelatie maakt de richting van de link niet uit; als A over hetzelfde gaat als B, dan gaat B ook over hetzelfde als A. Zoekwoorden die in pagina A voorkomen vormen tekstueel bewijsmateriaal dat pagina A relevant is voor de zoekvraag. Als de zoekwoorden vaak voorkomen in zowel pagina A als pagina B, dan is een link van A naar B een signaal dat het tekstuele bewijs voor A ook bewijs is voor de relevantie van B. Andersom zegt het tekstuele bewijs voor B ook iets over de relevantie van A. Het aantal lokale links tussen pagina A en andere lokale pagina's geeft aan hoeveel bewijs er voor A is. Hoe meer van die lokale links pagina A heeft, hoe meer bewijs er is voor de relevantie van A. De pagina met de meeste lokale links wordt beschouwd als het meest relevant.

Hoewel links een signaal geven dat twee gelinkte pagina's inhoudelijk aan elkaar gelateerd zijn, is die inhoudelijke relatie niet altijd even sterk. Om te meten hoe sterk twee pagina's inhoudelijk aan elkaar gerelateerd zijn gebruiken we de categie-informatie in Wikipedia. In Wikipedia blijkt de waarde van linkinformatie voor inhoudelijke relevantie afhankelijk te zijn van de relatie tussen twee gelinkte pagina's.

Een link tussen twee pagina's die over hetzelfde onderwerp gaan is effectiever bewijs voor de inhoudelijke relevantie van die pagina's dan een link tussen twee pagina's die over verschillende onderwerpen gaan. Deze bevinding bevestigd dat in Wikipedia linkinformatie als bewijs kan dienen voor de inhoudelijke relevantie van een pagina.

Uit deze bevindingen worden een aantal conclusies getrokken. Globale linkinformatie is onafhankelijk van de zoekvraag en geeft bewijs voor belangrijkheid (populariteit, authoriteit) maar niet voor inhoudelijke relevantie. Lokale linkinformatie is wel afhankelijk van de zoekvraag, en geeft bewijs voor inhoudelijke relevantie. Voor globale linkinformatie is de richting bepalend voor de betekenis ervan. Voor lokale linkfinformatie is de richting van minder belang.

Verdere ondersteuning voor deze conclusies vinden we in Wikipedia, waarbij de relatie tussen lokale linkinformatie en inhoudelijke relevantie verder verduidelijkt wordt. Ten eerste blijkt de hoeveelheid aan lokaal linkbewijs gerelateerd aan de hoeveelheid relevante informatie in een document, ongeacht de richting van de links. Ten tweede blijkt de fractie van het globale aantal links dat aanwezig is in de lokale linkstructuur (ongeacht de richting van de links) gerelateerd aan hoe specifiek het document over het zoekonderwerp gaat.

Hiermee is het antwoord op de vraag wat de waarde van linkinformatie voor het ordenen van zoekresultaten is, duidelijker en vollediger geworden. Linkinformatie kan bewijs vormen voor zowel populariteit als inhoudelijke relevantie. De betekenis van linkinformatie wordt bepaald door de richting van de links, de inhoudelijke relatie tussen de gelinkte documenten, en de selectie van links die worden gebruikt als bewijs.