Soekenjindekking van Suid-Arikaanse en Afrikaanse Webruimtes

Search engine coverage of South African and Afrikaans websites

Johan Breytenbach & Theo McDonald
Departement Rekenaarwetenskap en Informatika,
Universiteit van die Vrystaat, Bloemfontein
breytenbachj@gmail.com

Johan Breytenbach

Theo McDonald

Johan Breytenbach behaal die grade BCom (Wiskunde) en BComHons (Rekenaarwetenskap) aan die Universiteit van Stellenbosch en die Universiteit van die Vrystaat in 2004 en 2006 onderskeidelik. In 2009 voltooi hy sy Meestersgraad in Rekenaarwetenskap en Informatika aan die Universiteit van die Vrystaat. Sedert 2007 werk hy as programmeerder, dataverwerker en lektor in die Kaap. Sy navorsingsbelangstelling dek tans Internettegnologie en die Kennisekonomie.	**Johan Breytenbach** obtained the degrees BCom (Mathematics) and BComHons (Computer Science) from the Universities of Stellenbosch and the Free State in 2004 and 2006 respectively. He completed the Masters degree in Computer Science and Informatics at the University of the Free State in 2009. Johan has been employed as a programmer, data processor and lector in Cape Town since 2007. His research interests include Internet technology and the Economy of Knowledge.
Theo McDonald behaal die grade BSc en BScHons aan die Universiteit van die Vrystaat en die Meestersgraad in Wiskundige Statisitiek aan die Randse Afrikaanse Universiteit. Vanaf 1970-1972 werk hy as statistiese navorser by die Mediese Navorsingsraad in Pretoria en vanaf 1973-1977 as hoofwaardeerder by die Stadsraad van Pretoria. In 1977 verwerf hy die PhD graad in Wiskundige Statistiek aan die Universiteit van die Vrystaat en in 1979 word hy aangestel as senior lektor in die Departement Rekenaarwetenskap aan hierdie universiteit. Hy word bevorder tot medeprofessor in1989 en tot volle professor in 1996. Sy navorsingsbelangstelling is tans die wyse waarop Internetgebruikers soektogte doen.	**Theo McDonald** holds the degrees BSc and BScHons from the Uiversity of the Free State; and a Masters degree in Mathematical Statistics from the Rand Afrikaans University (now the University of Johannesburg). He was employed as a statistics researcher from 1970-1972 at the Medical Research Council in Pretoria and as chief evaluator at the Municipality of Pretoria from 1973-1977. He obtained the PhD degree in Mathematical Statistics at the University of the Free State in 1977 and in 1979 he was appointed a senior lecturer in the Department of Computer Science at the same university. In 1989 he was promoted to associate professor and in 1989 to full professor at the Univerity of the Free State. His research interest currently focuses on the manner in which Internet users are conducting searches on the Internet.

ABSTRACT

Search engine coverage of South African and Afrikaans websites

Search engines are web-based systems used for retrieving information from the Internet. They have economic power because of their positioning between information providers and information seekers. Search engines can influence the flow of information – possible business transactions – by the way information is indexed, stored, and portrayed as search results. If a search engine provides good coverage of website content from one group of information providers (grouped by country or language) to the detriment of another group, it will have economic implications for both groups. It is known that certain developed countries and the language(s) of these countries, have better coverage than other developed countries and their languages. This study investigates for the first time the website coverage of a developing country, South Africa, and one of its indigenous languages, Afrikaans.

What does the existence of search engine country bias and/or linguistic bias imply for developing countries such as South Africa? South African information providers would have reason for concern if information seekers' attention were continually routed abroad by biased search engines. South African information seekers would also be done an injustice if they cannot find local web content (or content in their indigenous languages) due to poor search engine coverage. Biased search results guide them away from cheaper, more convenient local information due to poor coverage of local content. Search engines are negatively impacted in turn, when users become tired of the poor local coverage and unwanted international search results and turn to other tools, such as local search engines, for information retrieval.

How severe are the effects of search engine bias on developing countries such as South Africa? The body of knowledge discussing search engine bias is very limited, and this study is motivated by stating that, given the possibility of negative economic implications of such bias for developing economies, more research on this topic is justified and urgently needed.

The study revealed that Western website content enjoys better coverage than South African website content. After further investigation it was also found that English website content enjoys better coverage than website content in Afrikaans. There is, therefore, a proven search engine bias in favour of Western developed countries and the English language.

Website visibility is also studied as a possible cause of search engine bias. It would seem plausible that a relationship may exist between the coverage of websites by search engines and how visible these websites are to a search engine's crawlers. For the determination of website visibility, the number of inlinks towards each sample domain was determined. This study shows that the higher visibility of websites from developed countries is a cause of search engine bias in favour of these websites.

With website visibility proven as a cause of bias, the study indicates that South Africa, and possibly other developing countries, is lagging far behind in the race to create highly visible websites surrounded by well-covered hyperlink structures – the kind of websites most likely to be covered by search engine crawlers. It is the responsibilty of information providers from developing countries to create more hyperlinks between websites from their countries, as well as creating visible website content in their indigenous languages.

Search engine coverage bias has negative economic implications for developing countries such as South Africa. This paper investigates the severity of country related coverage bias against websites from the South African domain(s) and the correlation between website coverage and website visibility. Other possible causes of coverage bias found in literature include indexing algorithms, ranking algorithms and lexicons struggling with non-English content. Information providers' lack of knowledge about website coverage and search engine tools is discussed as another possible cause of country bias.

KEY WORDS: Internet, Search engines, website coverage, coverage bias, language bias, website visibility, developing countries, Afrikaans website content

TREFWOORDE: Internet, soekenjin, websoektog, dekkingsydigheid, taalsydigheid, sigbaarheid van webtuiste, onwikkelende lande, Afrikaans

OPSOMMING

Soekenjins is webgebaseerde stelsels wat gebruik word om inligting vanaf die Web te verkry. As gevolg van hul posisie tussen inligtingsoekers en inligtingverskaffers beskik soekenjins oor ekonomiese mag. Soekenjins beïnvloed die vloei van inligting – moontlike besigheidstransaksies – deur die manier waarop inligting gedek, geïndekseer, gestoor en as soekresultate verskaf word. Indien ’n soekenjin goeie dekking verskaf aan ’n sekere groep (lande of tale) ten koste van ’n ander groep, sal dit ekonomiese gevolge vir beide groepe inhou. Inligtingsoekers, moontlike kliënte, word (meestal onwetend) gelei na die inligtingverskaffers wie se webinhoud die beste gedek word, deurdat hierdie webinhoud meer gereeld as relevante soekresultate aangedui word. Dit is bekend dat sekere ontwikkelde lande en hierdie lande se tale beter dekking geniet as ander ontwikkelde lande. Hierdie studie ondersoek vir die eerste keer die dekking van ’n ontwikkelende land, Suid-Afrika, en een van sy inheemse tale, Afrikaans. Daar is bevind dat Westerse (.com) webinhoud beter gedek word as Suid-Afrikaanse webinhoud en dat Engelse webinhoud beter dekking geniet as webinhoud in Afrikaans. Daar is gevolglik ’n sydigheid van soekenjins tot voordeel van Westerse (.com) ontwikkelde lande en die Engelse taal. Die sigbaarheid van webtuistes word ook bestudeer as ’n moontlike oorsaak van soekenjins se sydigheid.

INLEIDING

Volgens Comscore[1] word die gebruik van soekenjins steeds jaar-op-jaar meer gewild met die volume van soektogte wat aanhou styg gedurende 2006 en 2007. Soekenjins kan gesien word as die meeste Internetgebruikers se verkose middel vir inligtingsopsporing – 84% van Internetgebruikers dui aan dat hulle gereeld van soekenjins gebruik maak vir inligtingopsporing.[2] Anders gestel: soekenjins het hulself gevestig as die verkose koppelvlak tussen inligtingsoekers en inligtingverskaffers.

Soekenjinmaatskappye fokus daarop om webinhoud te versamel en hierdie webinhoud as soektogresultate aan inligtingsoekers te vertoon. Soekenjins vaar besigheidsgewys beter as meer inligtingsoekers van hul koppelvlak gebruik maak. Inligtingsoekers migreer natuurlik na die soekenjin wat die mees relevante inligting op die gemaklikste manier aan eindverbruikers bied. Inligtingsoekers sien akkurate, onsydige soekresultate meer relevant as sydige soekresultate[3] en migreer natuurlik na die onsydigste soekenjin wat tot hul beskikking is. Dit is dus in die beste belang van soekenjins, inligtingsoekers en inligtingverskaffers om beheerbare vorms van soekenjinsydigheid te minimaliseer.[3]

Vir die doeleindes van hierdie studie word soekenjinsydigheid gedefinieer as ’n onregverdige voordeel of nadeel wat deur soekenjins aan inligtingsoekers of inligtingverskaffers gegee word wanneer ’n soekenjin gebruik word om inligting vanaf die Web te verkry. ’n Land word benadeel deur soekenjinsydigheid indien ’n ondergemiddelde deel van daardie land se webtuistes geïndekseer is deur gewilde soekenjins. ’n Taal word benadeel deur soekenjintaalsydigheid indien ’n ondergemiddelde deel van die webinhoud in daardie taal gedek word deur gewilde soekenjins. Hierdie studie ondersoek vir die eerste keer of daar dekkingsydigheid ten opsigte van ’n ontwikkelende land, Suid-Afrika is, en of daar taalsydigheid ten opsigte van een van sy inheemse tale, Afrikaans, bestaan.

TEORETIESE AGTERGROND

In hierdie afdeling word bestaande literatuur rakende soekenjins en soekenjinsydigheid bespreek. Eers word daar gekyk na die interne werking van soekenjins en moontlike tegniese oorsake van soekenjins se taalsydigheid. 'n Bespreking van literatuur rakende sydigheid in soekenjins se plasingsalgoritmes volg, waarna literatuur rakende taalsydigheid, sydigheid teenoor sekere groepe gebruikers en dekkingsydigheid opgesom word.

Tegniese oorsake van soekenjinsydigheid

Die welbekende studie van Brin en Page,[4] wat die interne werking van die Google soekenjin beskryf, is gebruik as beginpunt in die soektog vir moontlike tegniese oorsake van soekenjinsydigheid. In die genoemde studie ontleed die outeurs 'n soekenjin se hoof funksionele komponente: 'n kuberkruiper en 'n URL (Uniform Resource Locator–webadres)-bediener, 'n stoorbediener en stoorplek, 'n indekseerder, 'n ankerstoorplek en 'n URL-ontleder, 'n leksikon, 'n plasingsalgoritme en 'n gebruikerskoppelvlak. Elkeen van hierdie komponente is vir hierdie studie se doeleindes ondersoek as 'n moontlike oorsaak van soekenjinsydigheid. Voordat daar verder op sydigheid gefokus word, volg 'n kort beskrywing van hierdie komponente.

Kuberkruipers, of spinnekoppe, is geoutomatiseerde programmatuur wat dele van die web stelselmatig deursoek. Kuberkruipers beweeg van webbladsy tot webbladsy deur hiperskakels te volg. Soekenjins gebruik kuberkruipers om webinhoud in te samel. Dit is die werk van 'n afsonderlike komponent, die URL-bediener, om te besluit watter webbladsye vervolgens deur die kuberkruipers besoek moet word. Verskillende algoritmes word hiervoor gebruik. Die kuberkruipers is verantwoordelik daarvoor om die webinhoud van al die besoekte webbladsye in te samel en dit te stoor in 'n stoorspasiebediener. Die stoorspasiebediener kompakteer die webinhoud en stoor dit ordelik in 'n stoorplek (repository). Die webinhoud word vanuit die stoorplek opgebreek in woorde en elke woord word, saam met inligting oor die woord se oorspronklike konteks, gestoor in 'n leksikon. Hierdie proses, waar webinhoud gestoor word in 'n geïndekseerde stoorplek en 'n leksikon, word algemeen na verwys as indeksering. 'n Komplekse soekenjinkomponent, die indekseerder, behartig die indeksering en maak gebruik van verskillende hutsalgoritmes om die leksikoninhoud met die oorspronklike brondokumente te verbind. Enige nuwe hiperskakels wat gedurende indeksering in die webinhoud gevind word, word gestoor in 'n hiperskakelstoor, verwerk deur 'n URL-ontleder, en uiteindelik na die URL-bediener gestuur sodat dit in die toekoms deur kuberkruipers besoek kan word.

'n Gebruikerskoppelvlak word deur inligtingsoekers gebruik om die geïndekseerde web-inhoud te deursoek deur sleutelwoorde in te tik. Indien die sleutelwoorde in die leksikon voorkom, kan die indekseerder die woord se oorspronklike konteks in die stoorplek vind en die oorspronklike webinhoud, of brondokumente, as 'n moontlike soektogresultaat voorstel. Die toepaslikheid van elke soektogresultaat, en dus die volgorde van die soektogresultate, word deur 'n plasingsalgoritme bepaal.

Die kwaliteit van soektogresultate hang af van die volledigheid van 'n soekenjin se webinhoud-stoorplek asook die kwaliteit van die interaksie tussen die leksikon, die indekseerder en die stoorplek. Indien 'n sekere webtuiste se inhoud nie volledig gedek is deur 'n kuberkruiper nie, sal daardie inhoud nie in die soekenjin se stoorplek wees nie. Die inligting wat onbekend is, sal dus nie as soektogresultate gewys kan word nie, want die soekenjin weet nie daarvan nie. Indien 'n webbladsy nie gereeld deur kuberkruipers besoek word nie, sal die stoorplek se webinhoud verouderd raak. Dit sal 'n negatiewe uitwerking op die kwaliteit van soektogresultate

hê. Buiten die probleme wat veroorsaak word deur onvolledige dekking en ongereelde besoeke deur kuberkruipers, is daar bewys dat soekenjins se kuberkruipers nie al die URLs wat aan die URL-bediener bekend is, deursoek nie. Wanneer 'n URL wel besoek word, word dit met wisselende vlakke van volledigheid gedek.[5]

Gee soekenjins voorkeur aan webtuistes vanuit sekere lande of aan webinhoud in sekere tale gedurende die tegniese prosesse van dekking en indeksering? Geen bewys van sydigheid kon gevind word in literatuur wat die tegniese werking van kuberkruipers bespreek nie. Die moontlikheid van sydigheid in die algoritmes wat URL- bedieners gebruik, word in enkele gevalle genoem, maar nie bewys nie. Die vraag rakende taalsydigheid in die tegniese werking van soekenjins plaas die fokus op die leksikon – die soekenjinkomponent wat alle aanvaarbare sleutelwoorde bevat. Groot soekenjins se leksikons heg 'n gewildheidstelling aan elke woord, wat op sy beurt gebruik word as 'n parameter in die soekenjin se plasingsalgoritme. Dit veroorsaak dat gewilde woorde meer geredelik as goedpassende soektogresultate aangedui word, en sodoende toeneem in gewildheid.[6] Hierdie gewildheidsparameter word in die volgende afdeling verder bespreek as 'n oorsaak van sydigheid teenoor minder algemene tale.

Nog 'n tegniese bron van taalsydigheid is die moontlikheid dat 'n soekenjin meer Engelse woorde in die leksikon bevat as woorde van enige ander taal. Literatuur rakende hierdie tipe sydigheid word later bespreek. Die verskil in sigbaarheid van webinhoud in verskillende tale word ook later in hierdie artikel bespreek.

Plasingsalgoritmes as 'n oorsaak van soekenjinsydigheid

Soekenjins verskaf inligtingsoekers met soektogresultate, georden van "mees toepaslik" tot "ontoepaslik". Soekenjins maak gebruik van 'n plasingsalgoritme om die rangorde van soektogresultate te bepaal. 'n Onsydige plasingsalgoritme is beter as 'n sydige plasingsalgoritme, aangesien inligtingsoekers migreer na die onsydigste soekenjin.[3] Wanneer soekenjinsydigheid bespreek word, word daar dikwels 'n beledigende vinger gewys in die rigting van plasingsalgoritmes,[7, 8] soms met goeie rede. Regsklagtes is al ingedien teen gewilde soekenjins omdat hul plasingsalgoritmes verdraaide of verwronge persepsies van maatskappye en markte skep.[9]

Die parameters van plasingsalgoritmes is ondersoek as moontlike oorsake van soekenjinsydigheid, en drie parameters is as moontlike oorsake vasgestel: (1) die struktuur en hoeveelheid hiperskakels rondom 'n webbladsy (sigbaarheid), (2) die gewildheidstelling van sleutelwoorde en webbladsye en (3) die gebruik van webtuistes se domeinname as 'n parameter in die plasingsalgoritme. Die struktuur van hiperskakels rondom 'n webbladsy kan daardie bladsy se dekking deur kuberkruipers beïnvloed,[5] asook daardie webbladsy se plasing as soektogresultaat.[8] Dit word duidelik gestel dat 'n groter hoeveelheid hiperskakels vanuit ander domeins, of in-skakels, 'n webbladsy se plasing positief sal beïnvloed.[7, 10, 11] Meer in-skakels na 'n webbladsy beteken dus beter dekking vir daardie bladsy, asook 'n beter plasing op die soektogresultaatranglys. Daar is ook bewys dat die gebruik van 'n sleutelwoord of 'n brondokument se gewildheidstelling as 'n plasingsparameter 'n sydige effek op soektogresultate kan hê.[6] Dit is egter moeilik om die sydige invloed van elkeen van die genoemde parameters te meet, aangesien so 'n eksperiment sou vereis dat een parameter verander terwyl al die ander konstant bly. Een moontlike ompad is om te aanvaar dat huidige algoritmes sydig is en liewer die aandag te fokus op die ontwikkeling van nuwe plasingsalgoritmes wat onsydig is vir enige waardes van hiperskakel-struktuur, gewildheid of domeinnaam. Onsydige plasingsalgoritmes is reeds voorgestel in literatuur, byvoorbeeld Cho en Adams,[12] en val buite die bestek van hierdie artikel.

Die kwaliteit van soekenjins en soektogresultate

Lewandowksi en Höchstötter[13] noem soekenjindekking – hoe deeglik 'n soekenjin se kuberkruipers 'n webtuiste dek – as een van die sleutelaanwysers van 'n soekenjin se kwaliteit. Die kwaliteit van beide die taaldekking en fisiese dekking word kortliks bespreek. 'n Soekenjin moet 'n redelike deel van alle domeins wat aan die soekenjin bekend is, volledig dek, en 'n redelike persentasie van elke taal se woorde in die leksikon vervat, om as 'n goeie soekenjin geklassifiseer te word. Nie-Engelse gebruikers moet hulself nie in 'n situasie bevind waar dit moeilik is om webinhoud in hul eie (inheemse) taal te bekom nie. Indien so 'n situasie deur soekenjins veroorsaak word, bestaan daar dekkingsydigheid en/of taalsydigheid wat die genoemde negatiewe ekonomiese gevolge vir die nie-Engelse soekenjingebruikers van 'n land sal inhou.

Daar is veelvuldige verwysings in bestaande literatuur gevind wat daarop dui dat soekenjins sydig ten gunste van Engelse webinhoud ten koste van webinhoud in enige ander taal staan. Meeste van hierdie studies noem dekking van nie-Engelse webinhoud as 'n sleutelaanwyser van 'n soekenjin se kwaliteit. Slegs een derde van alle webtuistes is nie Engels nie,[14] terwyl bykans twee derdes van alle Internetgebruikers nie Engelssprekend is nie.[15] Hierdie skewe verdeling skep 'n situasie waar nie-Engelse gebruikers kan sukkel om relevante inligting te vind wanneer hulle soekenjins gebruik. Lazarinis[16] het die kwaliteit van soektogresultate vir Griekse sleutelwoorde ondersoek en noem die swak dekking van Griekse webinhoud as 'n oorsaak van dekkingsydigheid. Efthimiadis[17] ondersteun hierdie stelling met verdere werk rakende die dekking van Griekse webinhoud. Bar-Ilan en Gutman[14] bestudeer die kwaliteit van Russiese, Franse, Hongaarse en Hebreeuse soektogresultate en som hul bevindings op deur te noem dat soekenjins wanaangepas is om inheemse tale te indekseer en in hul leksikons te bevat. Dit skep volgens die genoemde outeurs 'n situasie waar die meeste nie-Engelse webinhoud bloot "verlore raak in die kuberruim".

Die literatuur wat tot dusver bespreek is, lei tot enkele sinvolle gevolgtrekkings rakende die wisselwerking tussen dekking en taalsydigheid. Eerstens blyk dit dat daar minder hiperskakels voorkom tussen Engelse en nie-Engelse webtuistes as tussen webtuistes van dieselfde land of taal.[18] Hierdie situasie sal tot gevolg hê dat kuberkruipers sukkel om via hiperskakels op Engelse webbladsye by nie-Engelse webbladsye uit te kom – dus 'n oorsaak van taalsydigheid. Nog 'n moontlike wisselwerking tussen dekking en taalsydigheid bestaan omdat leksikons 'n groter persentasie van die Engelse woordeskat bevat as van enige ander taal se woordeskat.[16] Leksikons is minder sensitief vir verbuigings en streeksvorme van nie-Engelse woorde.[16] Hierdie situasie word deels veroorsaak deurdat Engelse webtuistes beter dekking geniet as nie-Engelse webtuistes.[5]

Soekenjin-bruikbaarheidsydigheid

Bruikbaarheid word gedefinieer as die eienskap van 'n rekenaarstelsel wat 'n gebruiker toelaat om die stelsel te gebruik vir die doel waarvoor dit ontwerp is.[19] Bruikbaarheid-sydigheid ontstaan wanneer 'n gebruiker nie die stelsel suksesvol kan gebruik nie as gevolg van wie die persoon is. Tradisioneel word ouderdom, geslag, ras, taalvoorkeur en persoonlike ondervinding as moontlike bruikbaarheidsgrense beskou.

Is soekenjins sydig teenoor sekere gebruikers op grond van hul ouderdom, geslag, ras, taalvoorkeur of ondervindingsvlak? McDonald en Blignaut[20] ondersoek die soekenjinvaardigheid van studente uit verskillende rassegroepe in Suid-Afrika en vind, na 'n kort opleidingsperiode, geen verskil tussen die soekenjingebruik van verskillende rassegroepe nie.

McDonald en Blignaut[21] het ook die soekenjinvaardigheid van groepe studente met verskillende ervaringsvlakke en verskillende taalvoorkeure getoets. Daar is gevind dat, ongeag die taalvoorkeur van ’n student, onervare studente baie swakker soekenjinvaardigheid toon as ervare studente. Daar is ook bewys dat nie-Engelse, ervare studente baie swakker soekenjinvaardigheid toon as Engelse, ervare studente.[21] Met die bespreekte literatuur in gedagte is een moontlike gevolgtrekking dat die oorsaak van nie-Engelse,ervare studente se swak vertoning lê by die feit dat twee derdes van alle webinhoud Engels is, en dat dit ’n nie-Engelse student bloot langer neem om deur Engelse soektogresultate te werk.

Geen literatuur kon gevind word wat die soekenjinvaardigheid van mense met verskillende ouderdomme vergelyk nie.

Soekenjindekkingsydigheid

Daar is slegs een noemenswaardige artikel rakende soekenjins se dekkingsydigheid in bestaande literatuur gevind. Hierdie studie deur Vaughan en Thelwall[5] bespreek dekkingsydigheid teenoor Oosterse lande se webtuistes. Daar is bewys dat soekenjins se kuberkruipers Oorsterse webtuistes gemiddeld 54% dek. Westerse (.com-domeinnaam vir groot Westerse maatskappye wat hulle besigheid op die Internet doen) webtuistes word egter gemiddeld 83% volledig gedek. Al die webtuistes wat gebruik is, het Engelse webinhoud bevat. ’n Soortgelyke navorsingsmetodiek (as die een van Vaughan en Thelwall[5]) is vir hierdie artikel gebruik, maar daar word spesifiek klem gelê op die taal waarin webinhoud gepubliseer word.

Ter opsomming van hierdie literatuuroorsig is dit belangrik om te noem dat gewilde soekenjins bewus is van dekkingsydigheid en taalsydigheid. Meeste van die gewilde soekenjins is reeds besig om lokale gebruikers-“portale” te skep, sodat ’n gebruiker sy land- en taalvoorkeur kan spesifiseer. Hierdie artikel poog om bronne van taalsydigheid uit te wys waaraan inligtingsoekers en inligtingverskaffers iets kan doen en om riglyne voor te stel waarvolgens die negatiewe ekonomiese gevolge van soekenjinsydigheid geminimaliseer kan word. Laasgenoemde sal positiewe gevolge hê vir soekenjins, inligtingverskaffers en inligtingsoekers.

PROBLEEMSTELLING

Hierdie studie is gebaseer op die aanname dat onsydige soekenjins ’n beter diens aan inligtingverskaffers en inligtingsoekers kan bied, en dat onsydigheid dus ’n gewenste eienskap van enige soekenjin is.[5, 13, 3] Hiermee in gedagte word die probleemstelling vir hierdie artikel soos volg omskryf:

- Soekenjins se dekkingsydigheid in ontwikkelde lande is gedeeltelik bewys.[5] Ondersoek die graad van dekkingsydigheid teenoor ontwikkelende Afrika-lande, met Suid-Afrika as voorbeeld.
- Indien bewyse van dekkingsydigheid gevind word, ondersoek die intensiteit van taalsydigheid teenoor inheemse tale, met Afrikaans as voorbeeld.
- Ondersoek die sigbaarheid van webtuistes as ’n oorsaak van bogenoemde sydighede.

METODOLOGIE

Vaughan en Thelwall[5] stel ’n metode voor vir die meting van dekkingsydigheid. In ooreenstemming met die voorgestelde metode is ’n onafhanklike kuberkruiper geprogrammeer (in Python), en

gebruik om die webinhoud van ’n ewekansige steekproef Suid-Afrikaanse webtuistes te ondersoek. Om ’n waarlik ewekansige steekproef te verseker, is dit nodig om ’n volledige lys van alle Suid-Afrikaanse webtuistes beskikbaar te hê. So ’n lys bestaan wel, maar is nie algemeen beskikbaar nie, omdat dit moontlik vir verkeerde doeleindes gebruik kan word. As alternatief, en in ooreenstemming met die metodologie van Vaughan en Thelwall,[5] is moontlike domeinname lukraak gegenereer en elke bestaande domein gefiltreer volgens streng riglyne: die webtuiste mag nie onder konstruksie wees nie, nie gebruik word om die domeinnaam te verkoop nie, nie wagwoordbeskermd wees nie en die webtuiste mag ook nie self ’n soekenjin wees nie. Dié siftingsproses het uiteindelik 289 webtuistes opgelewer. Hierdie steekproef is dubbel die grootte van Vaughan en Thelwall[5] se steekproef en voldoende vir ’n vergelykende studie.

Vir elkeen van die webtuistes is belangrike inligting ingesamel, onder andere die aantal bladsye waaruit elke webtuiste bestaan en die taal van die webinhoud. Slegs Engelse (237) en Afrikaanse (52) steekproefwebtuistes is gebruik.

’n Dekkingspersentasie is vir elke webtuiste bepaal deur die totale aantal webbladsye wat deur die onafhanklike kuberkruiper gevind is, te vergelyk met die aantal bladsye wat aan gewilde soekenjins bekend is. Google, Yahoo! en Windows Live (MSN) is vir hierdie deel van die eksperiment gebruik. Die 289 webbladsye is toe opgedeel in twee taalgroepe, Engelse webtuistes en Afrikaanse webtuistes, en die verskil in dekkingspersentasie tussen die twee groepe is ondersoek.

Om die sigbaarheid van webtuistes as ’n moontlike rede vir sydigheid te bestudeer, is die totale aantal in-skakels vir elke steekproefwebtuiste bereken. Daar is van gekombineerde “linkdomain” en “site” opdragte in die Yahoo koppelvlak gebruik gemaak om die aantal in-skakels na elkeen van die 289 steekproefwebtuistes te bepaal. Vergelykende data kon nie met die Google en MSN koppelvlakke verkry word nie en dus is net Yahoo! se inligting vir hierdie deel van die studie gebruik. Die korrelasie tussen die aantal in-skakels en die dekking is gevolglik bereken om te bepaal of daar ’n verwantskap tussen die twee bestaan.

RESULTATE

Die genoemde eksperimente het aangetoon dat Suid-Afrikaanse webtuistes gemiddeld 47% gedek word (sien Tabel 1). ’n Eenrigting variansie-analise het aangetoon dat die dekking van die verskillende soekenjins betekenisvol verskil ($p=0,00$) en dat hierdie verskil veroorsaak word deur die lae dekking van MSM. Hierdie resultate stem goed ooreen met die bevindings van Vaughan en Thelwall (2004:8) dat Oosterse webtuistes: China (61%), Singapore (49%) en Taiwan (51%) gemiddeld gedek word deur Google, Alta Vista en AllTheWeb. Wanneer ons hierdie bevindinge vergelyk met die 83% gemiddelde dekking wat .com-webtuistes geniet,[5] is dit duidelik dat daar soekenjindekkingsydigheid bestaan ten koste van sekere lande se webtuistes, met Suid-Afrika wat in hierdie studie as voorbeeld dien.

TABEL 1: Dekking van Suid-Afrikaanse webtuistes

SOEKENJIN	DEKKING (%)
Yahoo!	52
Google	56
MSN	34
Gemiddeld	47

Nadat dekkingsydigheid teenoor Suid-Afrika se webtuistes bevestig is, is die invloed van webinhoud se taal in hierdie sydigheid ondersoek. Die webtuistes in die oorspronklike steekproef is opgedeel in twee groepe: Afrikaanse webtuistes (n=52) en Engelse webtuistes (n=237). Tabel 2 vertoon die resultate van hierdie ondersoek. Vir beide die Yahoo! en die Google soekenjins is die gemiddelde dekking van Engelse webtuistes betekenisvol hoër as vir Afrikaanse webtuistes. Ook in die geval as al die soekenjins saam gegroepeer word, is die gemiddelde dekking vir Engelse webtuistes hoër as vir Afrikaanse webtuistes (49% teenoor 38%). Indien die hipotese gestel word dat daar geen verskil is tussen die dekking van Afrikaanse en Engelse webtuistes nie, kan hierdie hipotese verwerp word met statistiese sekerheid ($p<0.01$). Daar is dus bewys dat Engelse webtuistes beter dekking geniet (in die Suid-Afrikaanse webdomein) as webtuistes met webinhoud in die inheemse Suid-Afrikaanse taal, Afrikaans. Om die resultate tot dusver op te som: 'n Engelse .com webtuiste word gemiddeld 83% gedek,[5] terwyl 'n Afrikaanse webtuiste in die Suid-Afrikaanse webdomein (.co.za; .org.za; .gov.za) gemiddeld slegs 38% gedek word – 'n beduidende verskil van 45%.

TABEL 2: Persentasiedekking van Afrikaanse en Engelse webtuistes

SOEKENJIN	AFRIKAANS	ENGELS	p-waarde
Yahoo!	40	52	0,00
Google	43	59	0,00
MSN	29	35	0,29
Gemiddeld	38	49	0,00

Met dekkingsydigheid en taalsydigheid teenoor ontwikkelende lande se webtuistes bewys, is die steekproefwebtuistes se sigbaarheid ondersoek as moontlike oorsaak van hierdie sydighede. Om die sigbaarheid van webtuistes te meet, is die totale aantal in-skakels vir elke steekproefwebtuiste bereken. Hoe meer skakels na 'n webtuiste wys, hoe groter word daardie webtuiste se sigbaarheid vir kuberkruipers. Dit maak sin dat daar moontlik 'n verband bestaan tussen 'n webtuiste se dekking en daardie webtuiste se sigbaarheid. Na die berekening van elke tuiste se geweegde aantal in-skakels is die korrelasie tussen "sigbaarheid" en "persentasie dekking" vir elke webtuiste bereken. Die korrelasie-koëffisiënte tussen dekking en sigbaarheid word vertoon in Tabel 3.

TABEL 3: Korrelasie-koëffisiënte tussen sigbaarheid en dekking

STEEKPROEF	N	KORRELASIE-KOËFFISIËNT	p-WAARDE
Suid-Afrikaanse webtuistes	289	0,19	$p<0.05$
Afrikaanse webtuistes	52	0,48	$p<0.05$
Engelse webtuistes	237	0,18	$p<0.05$

Indien die nulhipotese gestel word dat daar geen verband is tussen webtuistes se dekking en sigbaarheid nie, dan kan hierdie hipotese vir al drie gevalle verwerp word. Wanneer die korrelasie tussen dekking en sigbaarheid vir die groep Afrikaanse webtuistes en die groep Engelse webtuistes vergelyk word, is daar 'n duidelike verskil tussen die korrelasie-koëffisiënte vir hierdie twee groepe. Twee gevolgtrekkings word vanuit hierdie resultate gemaak: (a) Afrikaanse webtuistes

met beter sigbaarheid word beter gedek en (b) sigbaarheid het 'n veel groter invloed op die dekking van Afrikaanse webtuistes as op die dekking van Engelse webtuistes.

BESPREKING

Wat beteken die bestaan van taalsydigheid vir ontwikkelende lande soos Suid-Afrika? Suid-Afrikaanse inligtingsoekers word deur sydige soektogresultate na webtuistes buite die Suid-Afrikaanse webdomein verwys omdat Suid-Afrikaanse, en Afrikaanse, webtuistes swakker gedek is. Dit kan 'n uitvloei van moontlike besigheidstransaksies uit die Suid-Afrikaanse webdomein veroorsaak – die mark vir Suid-Afrikaanse inligtingverskaffers. Soekenjins word negatief beïnvloed deurdat Suid-Afrikaanse verbruikers moeg raak vir swak resultate wat nie sensitief is teenoor soektogte in Afrikaans nie, en dan eerder begin gebruik maak van ander middele, byvoorbeeld plaaslike Suid-Afrikaanse soekenjins, om inligting van 'n meer onsydige aard te bekom.

Sommige soekenjins, soos Google,[11] is bewus van die oorsake van soekenjinsydigheid (bv. in plasingsalgoritmes) en is besig om hul deel te doen om hierdie ongewenste gevolge van sydigheid te minimaliseer. Om hierdie proses aan te help, word inligtingverskaffers aangeraai om hul webtuistes vir dekking te registreer deur middel van sogenaamde "Webmaster"-koppelvlakke. Die verantwoordelikheid om goeie dekking te verseker word dus, deur middel van hierdie koppelvlakke, teruggeplaas in die hande van inligtingverskaffers. Soekenjins kan wel van hul kant af meer moeite doen om URL-bedieners se algoritmes op 'n meer deursigtigige wyse bekend te maak.

Gedurende die tydperk van hierdie studie was Google die enigste grootskaalse soekenjin wat 'n geografiese dekkingsdiens aan inligtingverskaffers gebied het. Wanneer 'n webtuiste vir dekking geregistreer word, kan die inligtingverskaffer ook aandui watter land (volgens domein) die webtuiste se teikenmark is. Inligting wat vir 'n spesifieke land bedoel is, kry dan 'n hoër plasing wanneer iemand uit daardie land die inligting soek. 'n Soortgelyke diens waarmee inligtingverskaffers taalvoorkeure kan spesifiseer, kan baie help om taalsydigheid te minimaliseer.

Soos vroeër gestel, dra inligtingverskaffers en soekenjins 'n gesamentlike verantwoordelikheid vir die dekking en sigbaarheid van hul webtuistes. Dit wil voorkom asof dit juis die verwaarlosing van hierdie verantwoordelikheid is, veral in ontwikkelende lande soos Suid-Afrika, wat die grootste oorsaak van soekenjinsydigheid is. Inligtingverskaffers weet bloot nie genoeg van soekenjins se hantering van verskillende tale en verskillende domeins om hierdie verantwoordelikheid gestand te doen nie. Ter oplossing: Suid-Afrikaanse inligtingverskaffers behoort seker te maak dat *reeds geïndekseerde* webtuistes, verkieslik webtuistes met verwante en hoëkwaliteitinligting, hiperskakels na nuwe (nie-geïndekseerde) webtuistes bevat. Dit geld veral vir Afrikaanse webtuistes waar dekking sterk afhanklik is van sigbaarheid (sien resultate in Tabel 3). Inligtingverskaffers behoort al hul webtuistes by so veel as moontlik van die gewilde alledaagse soekenjins te registreer, en hulle behoort van Google se geografiese-dekkingsdiens en soortgelyke dienste gebruik te maak.

Wat sigbaarheid betref, is dit nodig om in gedagte te hou dat "...inligtingverskaffers meestal skakels na webtuistes in diesefde mark en domein verkies bo webskakels na ander domeins."[18] Met die sigbaarheidsresultate in Tabel 3 in gedagte is dit duidelik dat Suid-Afrika, en moontlik ander ontwikkelende lande, agter is in die wedloop om sigbare webtuistes, omring deur goeie hiperskakelstrukture, te vestig. Dit is Suid-Afrikaanse inligtingverskaffers se verantwoordelikheid om meer sinvolle skakels tussen Suid-Afrikaanse webtuistes te skep en hierdie agterstand in te haal.

TOEKOMSTIGE NAVORSING

Hierdie artikel het gefokus op dekkingsydigheid en taalsydigheid in ontwikkelende lande, met Suid-Afrika en Afrikaans as voorbeelde. Die sigbaarheid van webtuistes is ook bespreek as 'n moontlike oorsaak van soekenjinsydigheid. Daar is ander moontlike oorsake van soekenjinsydigheid genoem in die literatuuroorsig, insluitend plasingsalgoritmes en gebruikersydigheid, maar 'n volledige analise van hierdie temas val buite die bestek van hierdie artikel.

Hierdie artikel gebruik slegs data rakende Suid-Afrika en Afrikaanse en Engelse webtuistes vanuit die Suid-Afrikaanse webdomein. Vergelykende studies vir ander ontwikkelende lande en ander Afrikatale word verlang. Dit sal ook interessant wees om die dekking van plaaslike soekenjins te vergelyk met dié van internasionale soekenjins.

OPSOMMING

Soekenjinsydigheid het 'n negatiewe impak op die ekonomieë van ontwikkelende lande soos Suid-Afrika. Hierdie artikel bespreek die intensiteit van hierdie negatiewe impak en bewys die bestaan van dekkingsydigheid en taalsydigheid teenoor Suid-Afrikaanse en Afrikaanse webtuistes. Vervolgens is daar ook gekyk na die invloed van die taal waarin webinhoud gepubliseer word op dekking, asook die invloed van sigbaarheid op dekking. Daar word bewys dat die taal van webinhoud wel 'n rol speel in die dekking van daardie webinhoud en dat 'n meer sigbare webtuiste beter dekking sal geniet. Dit wil voorkom asof die swak sigbaarheid van nie-Engelse webinhoud die grootste oorsaak is van die swak dekking van nie-Engelse webinhoud en moontlike oplossings hiervoor word bespreek.

VERWYSINGS

1. Comscore. (2007). Comscore Releases October U.S. Search Engine Rankings. http://www.comscore.com/press/release.asp?press=1908 [30 June 2008].
2. Fallows, D. & Mudd, R. (2004). The Popularity and Importance of Search Engines. http://www.comscore.com [30 June 2008].
3. Goldman, E. (2006). Search Engine Bias and the Demise of Search Engine Utopianism. *Yale Journal of Law & Technology*, 2005-2006. http://ssrn.com/abstract=893892 [30 August 2008].
4. Brin, S. & Page, L. (1998). The Anatomy of a Large Scale Hypertectual Web Search Engine. *Computer Networks and ISDN Systems*, 30(1-7):107-117.
5. Vaughan, L.& Thellwall, M. (2004). Search Engine Coverage Bias: Evidence and Possible Causes. *Information Proccessing and Management*, 40(4): 693-707.
6. Cho, J. & Roy, S. (2004). Impact of search engines on page popularity. In *Proceedings of the 13th international Conference on World Wide Web* (New York, NY, USA, May 17 - 20, 2004).
7. Rogers, I. (2002). The Google PageRank Algorithm and How It Works. http://www.ianrogers.net/google-page-rank/ [30 July 2008].
8. SearchEngineHonesty. (2007). Search Engines Mechanics – How They Work. http://www.searchenginehonesty.com/search_engine.html [30 July 2008].
9. Grimmelmann, J. (2007). The Structure of Search Engine Law. *New York Law School Research Paper Series 06/07* No. 23. http://papers.ssrn.com/sol3/papers.cfm?abstract_id=979568 [30 June 2008].
10. Craven, P. (2002). Google's PageRank Explained And How To Make The Most Of It. http://www.webworkshop.net/pagerank.html [30 June 2008].
11. Cutts, M. (2008). Simplified Explanation of PageRank. http://answers.google.com/answers/threadview/id/223807.html [30 July 2008].
12. Cho, J. & Adams, R.E. (2005). Page quality: In search of an unbiased web ranking. In: *Proceedings of the 2005 ACM SIGMOD international conference on Management of Data*, Baltimore, Maryland, February 2005, pp.551-562.

13. Lewandowski, D. & Höchstötter, N. (2007). Web Searching: A Quality Management Perspective. In Zimmer, M., Spink, A. (eds): *Web Search: Interdisciplinary perspectives*. Dordrecht, Springer, 2007.
14. Bar-Ilan, J. & Gutman, T. (2005). How do search engines respond to some non-English queries? *Journal of Information Science,* 31(1):13-28.
15. Internet World Stats. (2008). Internet world users by language. http://www.internetworldstats.com/stats7.htm [30 January2009].
16. Lazarinis, F. (2007). Web retrieval systems and the Greek Language: do they have an understanding? *Journal of Information Science*. 33(5):622-636.
17. Efthimiadis, E.N., Malevris, N., Kousaridas, A., Lepeniotou, A. & Loutas, N. (2008). An Evaluation on How Search Engines respond to Greek Language Queries. In: *Proceedings of the 41st Hawaii International Conference on System Sciences*, 2008.
18. Thelwall, M. (2002). Evidence for the existence of geographic trends in university web site interlinking. *Journal of Documentation*, 58(5):563-574.
19. Preece, J., Rogers, Y. & Sharp, H. (2002). *Interaction Design: beyond human-computer interaction*. John Wiley & Sons, Inc. (14-18).
20. McDonald, T. & Blignaut, P. (2008). The effect of cultural differences on the efficiency of searches on a university website. In: *Proceedings of the 11th International Conference on Human-Computer Interaction*, Las Vegas, Nevada, July 22-27, 2005.
21. McDonald, T. & Blignaut, P. (2007). Language Aspects of Web Searching: An African Perspective. In: *Proceedings of sixth international conference on cultural attitudes towards technology and communication 2008*, Nimes, France, June 21-24, 2008, pp. 216-229.