De EFFECTIVITEIT van BLOGZOEKMACHINES

Reinier Meijer
Studentnummer: 5614228
Datum afronding: 11 juli 2009
Scriptie Master Documentaire informatiewetenschap
Universiteit van Amsterdam
Scriptiebegeleider: Dr.ir. Jaap Kamps
Tweede lezer: Dr. Henk Voorbij

Inhoudsopgave

Samenvatting 4
Abstract (Samenvatting in Engels) 5
1 Inleiding 6
1.1 Inleiding 6
1.2 Probleemstelling 7
1.3 Onderzoeksopzet en verantwoording methode 7
2 Blog 11
2.1 Inleiding 11
2.2 Ontstaan van blogs 11
2.3 Definities blog in de literatuur 12
2.4 Opstellen eigen definitie 16
2.5 Conclusie 17
3 Theoretisch kader 18
3.1 Inleiding 18
3.2 Positie binnen de informatiewetenschap 18
3.3 Informatie 19
3.4 Gebruikers 22
3.5 Systeem 24
3.6 Conclusie 26
4 Informatiezoekgedrag 27
4.1 Inleiding 27
4.2 Zoekvragen in blogzoekmachines 27
4.3 Wijze van zoeken met blogzoekmachines 30
4.4 Conclusie 31
5 Functionaliteiten blogzoekmachines 32
5.1 Inleiding 32
5.2 Ontstaan en ontwikkelaars 32
5.4 Verschillen en overeenkomsten in functionaliteiten 34
5.4.1 Opmerkingen 35
5.4.2 Verschillen ontsluiting blog posts 37
5.5 Conclusie 37
6 Opzet experiment 39
6.1 Inleiding 39
6.2 Doel experiment 39
6.3 Motivering zoekmachines 40
6.4 Wijze van opstellen zoekvragen 40

6.5 Wijze van uitvoeren zoekvragen ... 42
6.6 Wijze van analyse resultaten ... 46
7 Resultaten experimenten ... 47
7.1 Inleiding ... 47
7.2 Resultaten deelexperiment 1 ... 47
7.2.1 Resultaten m.b.t. totale aantal zoekvragen ... 47
7.2.2 Resultaten m.b.t. categorieën ... 48
7.3 Resultaten deelexperiment 2 ... 50
7.3.1 Resultaten m.b.t. totale aantal zoekvragen ... 50
7.3.2 Resultaten m.b.t. categorieën ... 50
7.4 Resultaten deelexperimenten 3a en 3b ... 52
7.4.1 Resultaten deelexperiment 3a ... 52
7.4.2 Resultaten deelexperiment 3b ... 53
7.5 Kwaliteit rangschikking ... 54
7.6 Statistieken alle deelexperimenten ... 57
7.7 Conclusie ... 57
8 Conclusie ... 59
Appendix A: Zoekvragen met indeling in categorieën ... 62
Appendix B: Data deelexperimenten ... 63
Literatuurlijst ... 64

SAMENVATTING

Miljoenen mensen besteden hun tijd aan het bloggen en het lezen van blogs, websites die in omgekeerd chronologische volgorde de bijdragen van één of meer individuen vermelden. Blogs vormen daarom voor velen belangrijke bronnen van informatie. Mede door middel van blogzoekmachines worden blogs en blog posts (bijdragen op blogs) ontsloten. In de wetenschappelijke literatuur kan echter nauwelijks onderzoek worden gevonden naar de effectiviteit van blogzoekmachines. Door middel van een experiment wordt onderzocht hoe effectief blogzoekmachines zijn bij het vinden van relevante blog posts en relevante overige webpagina's. Tevens wordt onderzocht wat de toegevoegde waarde van blogzoekmachines ten opzichte van algemene internetzoekmachines is. In het experiment worden vijftig zoekvragen in de blogzoekmachines Google Blog Search (NL), Technorati, Icerocket Blog Search en Twingly (spam-free) en in de algemene internetzoekmachine Google Search (NL) uitgevoerd.

Geconcludeerd kan worden dat gemiddeld 1/3 tot 1/2 van de resultaten in de top 10 van een blogzoekmachine relevant is, indien slechts blog posts op relevantie worden beoordeeld (en andere webpagina's dus als niet relevant worden beschouwd). Afhankelijk van de zoekmachine beschikt een top 10 resultatenpagina gemiddeld in 60 tot 90 procent van de zoekvragen over ten minste één relevant resultaat. Onderling vergeleken levert Twingly de meeste relevante resultaten.
Indien alle zoekmachineresultaten op relevantie worden beoordeeld, dan is gemiddeld 2/5 tot 7/10 van de resultaten in de top 10 van een blogzoekmachine relevant. Afhankelijk van de zoekmachine beschikt een top 10 resultatenpagina gemiddeld in 70 tot 100 procent van de zoekvragen over ten minste één relevant resultaat. Onderling vergeleken levert Google Blog Search de meeste relevante resultaten.
Tot slot kan worden geconcludeerd dat de toegevoegde waarde van blogzoekmachines afhankelijk is van de vraag of de internetter naast blog posts ook andere webpagina’s die in de informatiebehoefte voorzien, getoond wil krijgen. Zo ja, dan is de toegevoegde waarde beperkt. Zo nee, dan hebben sommige blogzoekmachines duidelijk toegevoegde waarde ten opzichte van een algemene internetzoekmachine zoals Google Search. Van de resultaten die Google Search toont is in dit geval minder dan één resultaat binnen de top 10 relevant. Ook beschikt een top 10 resultatenpagina gemiddeld in 1/3 van de zoekvragen over ten minste één relevant resultaat.

ABSTRACT (SAMENVATTING IN ENGELS)

Millions of people spend their time in blogging and reading blogs, websites that display, in reverse chronological order, the entries by one or more individuals. So, blogs are important sources of information for many people. Blogs and blog posts (the entries on blogs) can, among others, be retrieved via blog search engines. However, in the scientific literature, hardly any studies on the effectiveness of blog search engines can be found. In an experiment it is analysed how effective blog search engines are for retrieving relevant blog posts and relevant other web pages. The added value of blog search engines with respect to general internet search engines is analysed too. In the experiment fifty queries are submitted to blog search engines Google Blog Search (Dutch), Technorati, Icerocket Blog Search and Twingly (spam-free), and general internet search engines Google Search (Dutch).

Concluding, it can be said that on average 1/3 to 1/2 of the results in the top 10 of a blog search engine are relevant, when only blog posts are assessed on relevance (and other web pages are considered as not relevant). Depending on the blog search engine, a top 10 results page on average contains in 60 to 90 percent of the queries at least one relevant result. Compared with each other Twingly displays the most relevant results in the top 10.

When all search engine results are assessed on relevance, then on average 2/5 to 7/10 of the results in the top 10 of a blog search engine are relevant. Depending on the blog search engine, a top 10 results page on average contains in 70 to 100 percent of the queries at least one relevant result. Compared with each other Google Blog Search displays the most relevant results in the top 10.

Finally, it can be concluded that the added value of blog search engines depends on the following question: does the internet user want to find blog posts only, or does the internet user want to find other web pages that provide in the information need too? In the former case, the added value is limited. Otherwise, some blog search engines clearly have added value with respect to a general internet search engine like Google Search. In this case, on average less than one result of the results Google Search displays in the top 10 is relevant. Furthermore, on average a top 10 results page in Google Search contains in 1/3 of the queries at least one relevant result.

1 INLEIDING

1.1 Inleiding

Miljoenen mensen houden zich bezig met het bloggen en het lezen van blogs, websites die in omgekeerd chronologische volgorde de bijdragen van één of meer individuen vermelden. Op basis van marktonderzoek kan worden geconcludeerd dat wereldwijd naar schatting 184 miljoen personen ooit een blog zijn gestart.[1] Hiervan vormen de Chinezen, met 42,5 miljoen personen die een blog zijn gestart, de grootste bevolkingsgroep. In Nederland zijn naar schatting 1,7 miljoen personen ooit een blog gestart. Uit een onderzoek van Technorati blijkt daarnaast dat ongeveer 1,1% van de bloggers wekelijks of vaker blogt.[2] De rest van de bloggers blogt minder frequent. Veel bloggers publiceren persoonlijke verhalen op hun blogs.[3] Bloggers publiceren echter ook vaak meer inhoudelijke bijdragen over onderwerpen als technologie, politiek, reizen, muziek en nieuws in het algemeen. Marktonderzoek is niet alleen gedaan naar bloggers, maar ook naar lezers van blogs. 71% van de wereldwijde groep van 16- tot 54-jarigen die zich dagelijks of om de dag op het internet begeeft, leest wekelijks of vaker blogs.[4] Daarom kan worden gesteld dat blogs voor een zeer groot deel van de bevolking belangrijke bronnen van informatie vormen.

Voordat blogs kunnen worden geraadpleegd, zullen de URLs van deze blogs en de hierop verschijnende blog posts echter eerst bekend moeten zijn. Blogs en blog posts worden ontsloten door zoekmachines, directories en andere vormen van URL-vermelding.[5] Alleen zoekmachines kunnen echter op (zeer) grote schaal blogs en blog posts ontsluiten. Daarbij richten blogzoekmachines zich specifiek op blogs. Uit de literatuur kan worden afgeleid dat naar de effectiviteit van blogzoekmachines nog nauwelijks onderzoek is verricht. Dergelijk onderzoek is echter wel gewenst. Allereerst kan met dergelijk onderzoek de kennis van blogzoekmachines en information retrieval binnen de wetenschap worden vergroot. Ten tweede heeft de praktijk baat bij onderzoek naar de effectiviteit van blogzoekmachines. Allereerst wil de niet-zakelijke internetter op de hoogte blijven van de ontwikkelingen in zijn interessegebieden. Daarnaast wil de zakelijke internetter lezen op welke wijze over zijn bedrijf, de concurrenten van het bedrijf, de markt waarin het bedrijf zich bevindt, en de overige interesses van het bedrijf wordt geblogd. Beiden zullen bij het bepalen van hun zoekstrategie het gebruik van blogzoekmachines moeten overwegen. Concluderend kan daarom worden gesteld dat iedereen groot belang heeft bij meer inzicht in de effectiviteit van blogzoekmachines.

---

[1] Universal McCann 2008.
[2] Technorati 2008
[3] Universal McCann 2008
[4] Universal McCann 2008
[5] Denk bij andere vormen van URL-vermelding bijvoorbeeld aan overzichten van links naar blogs binnen een blog (‘blog rolls’) of een vermelding in een blog post.

1.2 Probleemstelling

Aldus kan de volgende probleemstelling worden opgesteld:

> Hoe effectief zijn blogzoekmachines?

1.3 Onderzoeksopzet en verantwoording methode

Na dit inleidende hoofdstuk, wordt in hoofdstuk 2 allereerst het begrip blog gedefinieerd. Dit onderzoek vindt plaats aan de hand van de volgende deelvraag:

> Wat wordt verstaan onder het begrip blog?

Het onderzoek naar deze deelvraag vindt plaats door middel van literatuurstudie. De literatuur wordt doorzocht in LISA: Library and Information Science Abstracts[6] (hierna: LISA), Library, Information Science & Technology Abstracts[7] (hierna: LISTA) en Google Scholar[8]. Aangezien de inhoud van ACM Digital Library[9] veelal technisch van aard is, wordt deze database voor deze deelvraag niet doorzocht. In de databases is gezocht op de termen ‘blog’, ‘blogs’, ‘weblog’, ‘weblogs’, ‘blogosphere’ in samenhang met de termen ‘content’ en ‘definition’ en ‘defining’. Indien dit weinig resultaten opleverde is tevens gezocht zonder de termen ‘content’ en ‘definition’.

Vervolgens richt hoofdstuk 3 zich op de positie die de ontsluiting van blogs inneemt binnen de informatiewetenschap. Evenals kaartenbakken in de negentiende en twintigste eeuw kunnen binnen hedendaagse ontsluitingssystemen, zoals internetzoekmachines, drie elementen worden onderkend. Deze drie elementen zijn informatie, gebruikers, en een systeem dat de informatiebehoeften van gebruikers en de beschikbare informatie matcht. Voor alle drie de elementen moet worden onderzocht waarin de blog een bijzondere positie ten opzichte van de standaard webpagina inneemt. Het onderzoek vindt plaats aan de hand van de volgende deelvraag:

> Binnen welk theoretisch kader ligt de ontsluiting van informatie in het algemeen, en de ontsluiting van blogs in het bijzonder?

Het onderzoek naar deze deelvraag wordt verricht door middel van literatuurstudie, aangezien de ontsluiting van informatie in het algemeen, en de ontsluiting van blogs in het bijzonder, al door een

---

[6] <http://www.csa.com/>
[7] <http://web.ebscohost.com/>
[8] <http://scholar.google.nl/>
[9] <http://portal.acm.org/dl.cfm>

aantal auteurs is bestudeerd. Opnieuw wordt de wetenschappelijke literatuur doorzocht in de databases LISA: Library and Information Science Abstracts (hierna: LISA), Library, Information Science & Technology Abstracts (hierna: LISTA) en Google Scholar. Tevens wordt de database ACM Digital Library doorzocht, aangezien de ontsluiting van blogs technisch van aard is. Waar de wetenschappelijke literatuur ontbreekt, wordt aanvullend de niet-wetenschappelijke literatuur doorzocht. Meer specifiek worden blogs doorzocht. Dit gebeurt met de volgende zoekmachines: Google Search[10], Google Blog Search[11] en Twingly[12] In de databases en internetzoekmachines wordt gezocht met termen zoals 'information seeking model', 'information seeking', 'information retrieval', 'web queries', 'blogs', 'blog search', 'blogs search' en 'framework'.

Na hoofdstuk 3 worden twee elementen nader onderzocht: het informatiezoekgedrag van internetters die zoeken met blogzoekmachines (hoofdstuk 4) en de functionaliteiten van zoekmachines (hoofdstuk 5).

In hoofdstuk 4 zal allereerst worden ingezoomd op de mensen die naar blogs zoeken. Daarbij wordt nadrukkelijk ingegaan op de informatie waarnaar en de wijze waarop internetters met blogzoekmachines zoeken. Meer specifiek vindt dit onderzoek plaats aan de hand van de volgende deelvraag:

Waarnaar en hoe zoeken internetters met blogzoekmachines?

Deze vraag wordt onderzocht door middel van literatuurstudie, aangezien over het zoeken naar blogs (en informatie in het algemeen) enkele publicaties bestaan. Omdat in dit geval de vraagstelling in de meeste wetenschappelijke artikelen slechts beknopt wordt behandeld, wordt ook voor deze vraag aanvullend de niet-wetenschappelijke literatuur bestudeerd. Daarbij wordt met name onderzoek van marktonderzoeksbureaus bestudeerd. In dit hoofdstuk wordt de literatuur doorzocht in LISA, LISTA, ACM Digital Library, Google Scholar, Google Search, Google Blog Search en Twingly. De wetenschappelijke literatuur wordt doorzocht op de termen zoals 'blog search', 'blogs search', 'information needs' en 'seekers'. De niet-wetenschappelijke literatuur is doorzocht op de termen 'blog search' en blogs search'. Indien dit veel resultaten opleverde, zijn de resultaten verfijnd door tevens de namen van bestaande blogzoekmachines in te geven ('Google Blog Search', 'Technorati', 'Icerocket' en 'Twingly').

Het laatste hoofdstuk dat de fundamenten voorafgaand aan de kern van het onderzoek legt is hoofdstuk 5. In hoofdstuk 5 worden de functionaliteiten van de zoekmachines die in het experiment van hoofdstuk 6 en 7 worden betrokken onderzocht. Daarbij wordt tevens de achtergrond waartegen

[10] <http://www.google.nl/>
[11] <http://blogsearch.google.nl/>
[12] <http://www.twingly.com/>

de functionaliteiten van de zoekmachines zijn ontstaan belicht. De functionaliteiten van de volgende vier blogzoekmachines worden onderzocht: Google Blog Search, Technorati, Icerocket Blog Search en Twingly. Daarnaast worden de functionaliteiten van de algemene internetzoekmachine Google Search onderzocht, aangezien Google Search en Google Blog Search met elkaar worden vergeleken in hoofdstuk 6 en 7. In hoofdstuk 5 wordt de volgende vraag onderzocht:

> Wat zijn de verschillen en overeenkomsten tussen de functionaliteiten van blogzoekmachines onderling, en ten opzichte van Google Search?

Deze vraag wordt onderzocht door de websites van de te onderzoeken zoekmachines te bezoeken, de interfaces te bestuderen en enkele zoekacties uit te voeren. Aanvullend wordt de niet-wetenschappelijke literatuur op gelijke wijze als in hoofdstuk 4 bestudeerd. De wetenschappelijke literatuur wordt niet bestudeerd, aangezien deze literatuur te oud is om hierop de actuele staat van de blogzoekmachines te baseren.

Nu in hoofdstukken 2 tot en met 5 de fundamenten voor het onderzoek naar de effectiviteit van blogzoekmachines zijn gelegd, wordt in hoofdstuk 6 en 7 toegekomen aan de effectiviteit zelf. Aangezien blogs op dezelfde wijze als sommige andere websites door middel van RSS-feeds kenbaar maken wat de meest recente webpagina's zijn, is het mogelijk dat blogzoekmachines meer dan alleen blog posts vinden en toegankelijk maken. Sommige internetters zullen deze overige webpagina's als resultaten binnen blogzoekmachines waarderen, terwijl anderen liever slechts blog posts als resultaten binnen blogzoekmachines willen vinden. Daarnaast zullen sommige internetters nauwelijks of nooit blogzoekmachines gebruiken, terwijl andere internetters deze zoekmachines veelvuldig gebruiken. Daarom vindt het onderzoek naar de effectiviteit plaats aan de hand van de volgende deelvraag:

> Hoe effectief zijn blogzoekmachines bij het vinden van relevante blog posts en relevante overige webpagina’s, en wat is de toegevoegde waarde ten opzichte van algemene internetzoekmachines?

Deze vraag wordt onderzocht door middel van een experiment, omdat literatuurstudie tot onvoldoende recente publicaties leidt.[13] In hoofdstuk 6 wordt allereerst de opzet van het experiment beschreven. De vier blogzoekmachines Google Blog Search, Technorati, Icerocket Blog Search en Twingly zullen worden beoordeeld op de retrieval effectiviteit: welke blogzoekmachine levert de meeste relevante resultaten in de top 10? In hoofdstuk 7 worden de resultaten voor twee verschillende typen internetters gepresenteerd: de internetter die alleen naar ‘echte’ blogs zoekt en de internetter die ook naar andere webpagina’s (zoals pagina’s met krantenartikelen) zoekt. De blogzoekmachines worden niet alleen

[13] De literatuurstudie is verricht door in LISA, LISTA, ACM Digital Library en Google Scholar te zoeken op de termen ‘blog search’ en ‘blogs search’, en de hieruit voortkomende resultaten grondig te bestuderen.

onderling vergeleken. Ook wordt de effectiviteit van Google Blog Search met de effectiviteit van Google Search vergeleken.

Tot slot volgt in hoofdstuk 8 de conclusie.

2 BLOG

2.1 Inleiding

Blogs (of weblogs) zijn sinds het eind van de vorige eeuw een veelvoorkomend type website. Voordat begonnen kan worden met het onderzoek naar de ontsluiting van blogs, zal eerst moeten worden onderzocht wat onder de term blog moet worden verstaan. Bij dit onderzoek wordt uitgegaan van de volgende deelvraag: 'Wat wordt verstaan onder het begrip blog?' In paragraaf 2.2 wordt allereerst kort het ontstaan van blogs belicht. In paragraaf 2.3 worden vervolgens de definities van blog in de literatuur onderzocht. Hierna wordt in paragraaf 2.4 een eigen definitie van blog opgesteld. Paragraaf 2.5 sluit het hoofdstuk af met een conclusie.
Vooraf wordt opgemerkt dat in plaats van de term 'blog' in de literatuur tevens de term 'weblog' regelmatig wordt gehanteerd. De twee termen worden als synoniemen beschouwd.[14] Aangezien de term 'blog' in de loop der tijd meer in gebruik is geraakt, wordt in dit onderzoek van deze term uitgegaan. Om dezelfde reden wordt de persoon die de blog bijhoudt als een 'blogger' aangeduid.

2.2 Ontstaan van blogs

In de literatuur wordt gesteld dat Jorn Barger in 1997 als eerste de term weblog hanteerde.[15] Op zijn persoonlijke website, Robot Wisdom[16], hield Barger een overzicht bij van links naar andere, zijns inziens interessante pagina's op het World Wide Web. Deze links werden in omgekeerd chronologische volgorde weergegeven, zodat steeds de nieuwste links bovenaan werden vermeld. Op zijn site vermeldde Barger in de FAQ uit 1999 reeds dat in plaats van de term weblog soms de term blog wordt gehanteerd.[17] Ook benadrukte Barger dat het onderscheid tussen een weblog en een dagboek zijns inziens moeilijk te maken is. Bloggers plaatsten namelijk geregeld niet alleen links naar interessante pagina's, maar gaven ook meningen over de betreffende onderwerpen.

Voordat de term weblog werd gehanteerd, konden echter ook reeds in omgekeerd chronologische volgorde geplaatste linkoverzichten worden onderkend. De 'What's New'-pagina die Tim Berner-Lee in 1992 op http://info.cern.ch creëerde, zou daarom als eerste weblog kunnen worden beschouwd.[18]

Met de toename van het aantal internetaansluitingen en het verschijnen van eenvoudige weblogapplicaties, startten steeds meer mensen een weblog, en werd steeds vaker de term 'blog' in

[14] Walker 2005.
[15] Mishne 2007, p. 8 en Herring e.a. 2004, p. 1.
[16] <http://robotwisdom.com> (niet beschikbaar).
[17] <http://web.archive.org/web/20000817093828/www.robotwisdom.com/weblogs/>
[18] Mishne 2007, p. 16.

plaats van ‘weblog’ gebruikt.[19] Blog posts met beknopte beschrijvingen van webpagina’s in de beginjaren ontwikkelden zich tot blog posts met langere beschrijvingen, commentaren en reflecties.[20] Tegenwoordig komen vaak blog posts (de artikelen op blogs) voor die geen betrekking hebben op webpagina’s, maar waarin de blogger allerlei onderwerpen becommentarieert en zijn belevenissen beschrijft.

2.3 Definities blog in de literatuur

Sinds het eind van de jaren negentig van de vorige eeuw hebben bloggers en onderzoekers uiteenlopende definities van blog opgesteld.[21] In deze definities komen veelal elementen op het gebied van inhoud en/of structuur voor. Boyd merkt op dat onderzoekers aan het begin van hun artikel steeds definiëren wat een blog is. Hieruit maakt Boyd op dat er geen eenduidige, gangbare definities van blog onder de onderzoekers bestaan. Als gevolg hiervan baseren onderzoekers zich vooral op definities uit de technologische sector en uit woordenboeken. Hierna worden allereerst de belangrijkste definities uit woordenboeken en encyclopedieën weergegeven. Vervolgens komen definities van bloggers en onderzoekers aan bod.

Om allereerst te weten hoe een algemeen woordenboek de term blog definieert, kan de Merriam-Webster Online Dictionary wordt geraadpleegd. Dit woordenboek bevat de volgende definitie van blog:

> *“a Web site that contains an online personal journal with reflections, comments, and often hyperlinks provided by the writer”* [22]

In deze definitie wordt vooral het persoonlijke karakter van de blog benadrukt.

De Oxford English Dictionary Online hanteert een uitgebreidere definitie:

> *“A frequently updated web site consisting of personal observations, excerpts from other sources, etc., typically run by a single person, and usually with hyperlinks to other sites; an online journal or diary”* [23]

Deze definitie is ruimer doordat de inhoud van blogs mede met ‘etc.’ wordt omschreven. Blogs kunnen hiermee van alles bevatten. In tegenstelling tot de Merriam-Webster Online Dictionary noemt

[19] Mishne 2007, p. 8.
[20] Vgl. Herring e.a, p. 9.
[21] Boyd 2006.
[22] ‘Blog’ in Merriam-Webster Online Dictionary 2009.
[23] ‘Weblog’ (2) in Oxford English Dictionary Online 2009.

de Oxford English Dictionary Online in de definitie tevens het principe dat deze webpagina's veelvuldig worden bijgewerkt.

Vervolgens kan de definitie van blog ook in een vakgerichte encyclopedie, de Routledge Encyclopedia of Narrative Theory, worden opgezocht. Walker definieert de blog in deze encyclopedie als volgt:

> *"a frequently updated website consisting of dated entries arranged in reverse chronological order so the most recent post appears first"* [24]

Het persoonlijke element dat in de vorige twee definities redelijk nadrukkelijk wordt genoemd, ontbreekt in de definitie van de Walker. Slechts structurele elementen komen in de definitie tot uitdrukking.
Een dergelijke definitie waarin slechts structurele elementen van blogs voorkomen, wordt veelvuldig gehanteerd.[25]

Naast de definities in woordenboeken en encyclopedieën, kunnen ook de definities uit de technologische sector worden geraadpleegd.

Allereerst kan de definitie van Technorati worden geraadpleegd. Technorati is vooral bekend vanwege zijn blogzoekmachine, maar biedt daarnaast ook andere diensten aan om de '*global online conversation*' toegankelijk te maken. In een recente blog post meldt het internetbedrijf dat de definitie op Wikipedia dicht komt bij het idee van blog zoals dat leeft bij Technorati.[26] De (inmiddels licht bijgeschaafde) definitie op Wikipedia luidt:

> "*A blog (a contraction of the term weblog) is a type of website, usually maintained by an individual with regular entries of commentary, descriptions of events, or other material such as graphics or video. Entries are commonly displayed in reverse-chronological order.*" [27]

Doordat woorden als '*usually*' en '*commonly*' worden gebruikt, stelt deze definitie nauwelijks grenzen aan de blog. Technorati gaat vervolgens in op blogs van mainstream media. Denk hierbij bijvoorbeeld aan blogs van verslaggevers op krantensites. Ook deze blogs zijn volgens het bedrijf gewoon onderdeel van de blogosphere. Als de eigenaren deze blogs aanmelden, dan zal Technorati deze blogs gewoon toegankelijk maken.

---

[24] Walker 2005.
[25] Vgl. Herring e.a. 2004, Kumar e.a. 2004, p. 35 en Mishne 2007, p. 8.
[26] Technorati 2009.
[27] 'Blog' in Wikipedia 2009.

Tot slot kunnen de definities van onderzoekers worden geraadpleegd.

Allereerst kan daarbij de beschouwing van Boyd worden bestudeerd.[28] Boyd bekritiseert een groot aantal definities, om zich vervolgens op een drietal karakteristieken van blogs te richten. Allereerst vindt Boyd dat de blog zodanig verbonden is met de activiteit van het bloggen, dat in de definitie op dit laatste de nadruk moet liggen om tot een kritisch inzicht te kunnen komen. Daarnaast is de auteur van mening dat de blog, net als de radio, tv en krant als een medium moet worden beschouwd. Ook de blog is namelijk een verlengstuk waarmee je je naar anderen toe kunt uiten. Tot slot is de auteur van mening dat de grenzen van de blog niet worden gesteld door de technologie (meer specifiek de blogsoftware). De blog is namelijk een sociaal geconstrueerd medium. Als de technologie evolueert, dan veranderen naar haar mening de eigenschappen van de blog mee. Op basis van voorgaande analyse komt Boyd tot de volgende definitie van blog:

> "*a culture-driven medium upon which the practice of blogging can occur*"

De auteur stelt vervolgens vast dat deze definitie ruimte laat voor diversiteit van blogs qua structuur en inhoud.

Een tweede definitie van een onderzoeker is de definitie van Mishne. Mishne hanteert een eigen interpretatie van de definitie van Winer. In deze interpretatie is een blog:

> *"a web page voicing the unedited, unregulated voice of an individual; its structure and format are just means for expressing the individual's views in an easier, more appealing, more interactive manner"* [29]

In deze definitie staat de persoon achter de blog centraal, en is het principe dat blog posts in omgekeerd chronologische volgorde worden weergegeven niet duidelijk aanwezig.

Vervolgens kan ook de definitie van Wilde worden geraadpleegd. In een onderzoek naar de structuren van blogs en blog posts definieert Wilde blogging, blogs en blog posts als volgt:

> "*Blogging in its most general definition is a timed series of entries (blog posts) in HTML, and the collection of all these posts is referred to as the blog. A blog typically has a homepage (which shows the most recent posts), some kind of archive (making available older posts) and well-defined authorship, describing who is writing the blog.*" [30]

[28] Boyd 2006.
[29] Mishne 2007, p. 8.
[30] Wilde 2008, p. 403.

Wilde laat in deze definitie het onderscheid tussen blogs en blog posts duidelijk tot uitdrukking komen. Daarbij schenkt Wilde aandacht aan de karakteristieken van een blog: een blog heeft een homepage en een soort archief en vermeldt duidelijk wie de auteur van de blog is.

Naast de definities van Boyd, Mishne en Wilde, kan ook de definitie van Agarwal en Liu worden genoemd. Zij definiëren een blog als:

> *"a website that displays, in reverse chronological order, the entries by one or more individuals and usually has links to comments on specific postings"* [31]

Ten opzichte van de definitie van Walker bestaan aldus enkele verschillen. In de definitie van Agarwal en Liu ontbreekt de woordengroep '*frequently updated*', maar is wel aanwezig wie de blog maakt en wordt vermeld dat de website meestal links naar commentaren op specifieke posts bevat.

Tot slot kan de definitie van Lomborg worden genoemd. Lomborg vindt dat de sociale en communicatieve functies van de blog te weinig tot uitdrukking komen in definities zoals die van Agarwal en Liu. Vanuit socio-pragmatisch perspectief komt de auteur tot de volgende definitie van blog:

> "*an author–driven, asynchronous and informal genre of CMC* [computer-mediated communication; RM] *that uses various modalities and entails some interactivity*" [32]

Deze definitie stelde Lomborg op op basis van elementen die ze aantrof in blogs. Zo stelde ze vast dat blogs meestal één auteur hebben, een informele stijl hebben, asynchroon[33] en blijvend zijn, en interactieve onderdelen hebben. Lomborg geeft aan deze ruime definitie ruimte laat voor verdere ontwikkeling van de blog.

Bar-Ilan behandelt de definities van blog vanuit allerlei hoeken, maar komt er niet toe zelf een definitie op te stellen.[34] Volgens de auteur kunnen op allerlei gebieden (inhoud, formaat, auteurschap) verschillen tussen blogs worden onderkend. Aldus is het lastig voor blogs een definitie op te stellen.

---

[31] Agarwal & Liu 2008, p.18.
[32] Lomborg 2009.
[33] Auteur en lezer hoeven niet tegelijkertijd online te zijn om op elkaar te reageren.
[34] Bar-Ilan 2005, p. 297.

2.4 Opstellen eigen definitie

Indien de definities uit paragraaf 2.3 worden geanalyseerd, dan kan een onderscheid worden gemaakt tussen de elementen die mijns inziens niet in de definitie thuis horen, en de elementen die mijns inziens wel in de definitie behoren voor te komen.

Allereerst zal worden ingegaan op de elementen die mijns inziens niet in de definitie thuis horen. Eerste element betreft de zienswijze van Boyd dat de blog een medium is. Mijns inziens gaat het te ver om de blog als een medium te beschouwen. Op een blog kun je je op dezelfde manier uiten als op een standaard website of op een internetforum. Aangezien webpagina's zich meestal via elektronische netwerken verspreiden, is het geheel van elektronische netwerken als medium te beschouwen. Vervolgens worden via deze netwerken allerlei webpagina's en e-mails verspreid.

Daarnaast kunnen in een aantal definities elementen worden onderkend, die 'meestal' of 'typisch' een blog karakteriseren. Merriam-Webster Online Dictionary, Oxford English Dictionary Online en Agarwal en Liu geven bijvoorbeeld in hun definitie weer dat blogs vaak links bevatten. Ook geven Oxford English Dictionary Online en Wikipedia in hun definitie weer dat blogs typisch door één persoon worden bijgehouden. Omdat deze elementen niet in elke blog tot uiting komen, behoren deze elementen mijns inziens niet in de definitie voor te komen.

Tot slot komt in een aantal definities tot uitdrukking dat blogs veelvuldig worden bijgewerkt.[35] Mijns inziens kenmerkt dit niet alle blogs. Een groot aantal blogs wordt nauwelijks bijgewerkt, maar kan op basis van de overige kenmerken toch als blog worden beschouwd.

Vervolgens kunnen de elementen worden onderkend die mijns inziens wel in de definitie behoren voor te komen. Allereerst moet naar mijn mening overeenkomstig de definitie in Wikipedia de blog als een type website worden beschouwd. Zoals hiervoor genoemd bestaan blogs namelijk naast standaard websites en internetfora. De blog als website strekt zich vrijwel altijd uit over meerdere webpagina's.[36]

Een ander element dat mijns inziens in de definitie hoort voor te komen is het element dat de blog posts binnen een blog op omgekeerd chronologische volgorde worden geordend. Dit element komt niet in alle bovengenoemde definities voor, maar mijns inziens bepaalt dit element zodanig of een website een blog is, dat het in de definitie moet worden opgenomen.

[35] Vgl. 'Weblog' (2) in Oxford English Dictionary Online 2009; Walker 2005.
[36] In de toekomst zouden sommige blogs zich overigens tot één pagina met één URL (Uniform Resource Locator; het adres van een webpagina) kunnen beperken met de verdere opkomst van 'asynchronous JavaScript and XML' (Ajax).

Aldus wordt de volgende definitie van blog opgesteld:

> “een type website dat in omgekeerd chronologische volgorde de bijdragen van één of meer individuen vermeldt”

Opgemerkt dient te worden dat voor het experiment dat in hoofdstuk 6 wordt beschreven, op een vereenvoudigde wijze wordt getoetst of een webpagina een blog post is. In het experiment wordt een webpagina niet als blog post beschouwd wanneer de webpagina geen reacties c.q. trackback links toont én geen mogelijkheid biedt tot het plaatsen van reacties c.q. trackback links. Op basis van deze regel worden spam en nieuwsartikelen van traditionele media reeds grotendeels uitgesloten. Aanvullend wordt een webpagina niet als blog post beschouwd indien de webpagina een artikel van een persbureau, krant, tijdschrift, tv-station of radiozender bevat zonder herkenbare toevoeging of aanpassing hiervan. Tot slot wordt een webpagina niet als blog post beschouwd indien de webpagina de lay-out van een forum topic of forumoverzichtspagina heeft én tevens de term forum (of een synoniem hiervan in welke taal dan ook) vermeldt. Op deze wijze kan eenvoudig zo objectief mogelijk worden getoetst of sprake is van een blog post.

2.5 Conclusie

De deelvraag ‘Wat wordt verstaan onder het begrip blog?’ kan aldus als volgt worden beantwoord. In de literatuur bestaat geen eenduidige definitie van blog. In een aantal definities komen elementen voor die niet in alle blogs aanwezig zijn. Deze elementen behoren mijns inziens niet in de definitie voor te komen. Mijns inziens moet als blog worden beschouwd: het type website dat in omgekeerd chronologische volgorde de bijdragen van één of meer individuen vermeldt. Om blog posts, de bijdragen op blogs, te identificeren voor het experiment van hoofdstuk 6 en 7 worden aanvullende eisen gesteld, om eenvoudig zo objectief mogelijk vast te stellen of sprake is van een blog post.

3 THEORETISCH KADER

3.1 Inleiding

In het vorige hoofdstuk is duidelijk geworden dat een blog een type website is dat in omgekeerd chronologische volgorde de bijdragen van één of meer individuen vermeldt. Vervolgens is de vraag hoe de ontsluiting van blogs zich verhoudt tot de ontsluiting van informatie in het algemeen. Dit onderzoek wordt verricht aan de hand van de volgende deelvraag: 'Binnen welk theoretisch kader ligt de ontsluiting van informatie in het algemeen, en de ontsluiting van blogs in het bijzonder?' Om deze deelvraag te kunnen beantwoorden wordt in paragraaf 3.2 allereerst onderzocht welke positie de ontsluiting van blogs binnen de informatiewetenschap inneemt. Vervolgens wordt in de drie daaropvolgende paragrafen afzonderlijk ingegaan op de drie elementen die binnen een ontsluitingssysteem een rol spelen:

- een verzameling informatie (paragraaf 3.3)
- de gebruikers (paragraaf 3.4)
- een systeem dat de informatiebehoeften van gebruikers en de verzameling informatie matcht (paragraaf 3.5)

Voor ieder onderdeel wordt onderzocht waarin het betreffende onderdeel tot verschillen tussen blogs en standaard webpagina's leidt. Paragraaf 3.6 sluit het hoofdstuk af met een conclusie. Aan het eind van deze conclusie wordt tevens vastgesteld op welke aspecten nader moet worden ingegaan, alvorens het experiment naar de effectiviteit van blogzoekmachines te kunnen uitvoeren.

3.2 Positie binnen de informatiewetenschap

De informatiewetenschap is een interdisciplinair onderzoeksgebied dat zich zowel theoretisch als op toegepaste wijze bezig houdt met het verzamelen, classificeren, opslaan, toegankelijk maken en verspreiden van opgeslagen kennis.[37] Op dezelfde wijze als in de negentiende en twintigste eeuw documenten werden ontsloten met behulp van kaartenbakken, worden tegenwoordig webpagina's op internet ontsloten met behulp van zoekmachines.[38] Afgezien van het verschil dat de ontsluiting destijds handmatig plaatsvond en deze tegenwoordig geautomatiseerd plaatsvindt, verschillen de kaartenbak en de zoekmachine nauwelijks.

[37] Vgl. 'Information science' in Merriam-Webster Online Dictionary 2009.
[38] Vgl. Rayward 1994.

Sinds het doorbreken van het internet in de jaren negentig van de twintigste eeuw zijn steeds meer webpagina's (met opgeslagen kennis) op het internet gepubliceerd. Om deze webpagina's te kunnen bekijken, zal de URL van de webpagina bij de potentiële lezer bekend moeten zijn. De lezer kan de URL bijvoorbeeld van iemand horen of deze URL in een papieren publicatie lezen, maar meestal zal de lezer de URL in een e-mail of op een webpagina lezen of de URL (ongemerkt) aanklikken. De webpagina's die URL's van andere webpagina's vermelden kunnen daarbij in de volgende drie categorieën worden onderscheiden:

- de webpagina is een resultatenpagina van een zoekmachine,
- de webpagina is een overzichtspagina met links ('directory'),
- de webpagina is noch een resultatenpagina van een zoekmachine, noch een overzichtspagina met links.

Het onderhavige onderzoek bevindt zich op het gebied van het toegankelijk maken van webpagina's door zoekmachines. Webpagina's op het internet konden sinds halverwege de jaren negentig worden doorzocht met algemene internetzoekmachines zoals Lycos en AltaVista. Tegen het eind van de jaren negentig raakten nieuwe vormen van opgeslagen kennis, zoals blogs, steeds meer ingeburgerd op het internet. Omdat blog posts veelal op de actualiteit waren gebaseerd en internetzoekmachines er niet in slaagden deze posts onmiddellijk doorzoekbaar te maken, ontstond ruimte voor de ontwikkeling van zoekmachines die zich specifiek op de inhoud van blogs richten. Blogzoekmachines zoals Technorati deden hun intrede. Doordat algemene internetzoekmachines sindsdien steeds sneller en vaker webpagina's zijn gaan indexeren, kunnen blogzoekmachines en algemene internetzoekmachines op het gebied van blogs als concurrenten worden beschouwd.

In de drie volgende paragrafen wordt ingegaan op drie elementen die binnen een zoekmachine en ieder ander ontsluitingssysteem een rol spelen:

- informatie (paragraaf 3.3)
- gebruikers (paragraaf 3.4)
- een systeem dat de informatiebehoeften van gebruikers en de informatie matcht[39] (paragraaf 3.5)

3.3 Informatie

Het onderhavige onderzoek neemt als uitgangspunt dat een document informatie bevat[40], en dat een webpagina wordt beschouwd als een type document. Mishne benoemt vanuit het oogpunt van

[39] Chowdhury 2004, p.2.

informatietoegang op drie niveaus verschillen tussen de blogspace[41] en verzamelingen andersoortige documenten:

- inhoud
- structuur
- datum- en tijdvermelding[42]

Allereerst merkt Mishne over de inhoud van blogs op dat deze vaak van persoonlijke aard is. Blogs bevatten beschrijvingen van het leven en de omgeving van de blogger. Daarnaast bevatten blogs gedachten, emoties en commentaren ten aanzien van uiteenlopende onderwerpen. Mishne merkt op dat dit type inhoud nauwelijks beschikbaar is in andere publieke corpora. Alleen in blogs kan het op grote schaal worden gevonden. Mijns inziens moet ten aanzien van dit eerste verschil worden opgemerkt dat de inhoud van een groot aantal andere blogs niet of nauwelijks persoonlijk van aard is. Ten aanzien van de verhouding tussen deze blogs en de blogs van persoonlijke aard kan in de wetenschappelijk literatuur geen onderzoek worden gevonden.
Het tweede verschil tussen blogs en verzamelingen andersoortige documenten ligt op het gebied van structuur. Blogs vormen een dicht sociaal netwerk van mensen, met een overvloed aan gemeenschappen en een snelle verspreiding van informatie. Verzamelingen andersoortige documenten zijn veel minder met elkaar verbonden.
Het laatste verschil bevindt zich op het gebied van vastlegging van het tijdstip van publicatie. In blogs wordt iedere blog post, commentaar of trackback van een nauwkeurige datum- en tijdstempel voorzien, waardoor publicatiedatum en -tijd vaak tot op de minuut nauwkeurig kunnen worden bepaald. Verzamelingen andersoortige documenten vermelden meestal niet structureel de datum en tijd van publicatie.

Thelwall en Hasler hebben de blog met een specifieker type document vergeleken: de standaard webpagina.[43] Naast het reeds door Mishne genoemde verschil van datumvermelding, zien de auteurs dat de structuur van de individuele blog sterk verschilt van de standaard webpagina. De individuele blog bestaat uit een groot aantal blog posts die op tal van manieren aan de bezoeker van de blog kunnen worden gepresenteerd. Blog posts kunnen bijvoorbeeld afzonderlijk worden getoond, maar ook gegroepeerd per week of maand, of als een lijst op de homepage. Aldus kan dezelfde blog post op allerlei pagina's binnen de blog worden weergegeven. Thelwall en Hasler merken aanvullend op dat een aantal blogs over de mogelijkheid beschikt tot het aanbieden van blog posts in het Rich Site Summary (RSS) formaat. Omdat het RSS-formaat een gestandaardiseerd formaat is, kunnen systemen eenvoudiger met de blog posts overweg. In het RSS-formaat worden de recentste blog posts met hun

---

[40] Buckland 1991 en Buckland 1997.
[41] De verzameling blogs, tevens bekend onder de naam blogosphere.
[42] Mishne 2007, p. 2.
[43] Thelwall en Hasler 2007, p. 469.

volledige inhoud of een gedeelte hiervan weergegeven. Bij iedere blog post wordt een link geplaatst naar de pagina waarop de specifieke post in de originele opmaak kan worden gevonden.

Herring e.a. hebben een continuüm opgesteld waarin de positie van blogs tussen standaard webpagina's en asynchrone customer-generated communication (CMC)[44] wordt weergegeven.[45] De drie typen webpagina's kunnen volgens de auteurs op drie vlakken worden onderscheiden. Dit onderscheid wordt weergegeven in Tabel 1.

Tabel 1 Verschillen tussen typen webpagina's

Standaard webpagina's	Blogs	Asynchrone CMC
Nauwelijks bijgewerkt	Veelvuldig bijgewerkt	Constant bijgewerkt
Asymmetrische verspreiding	Asymmetrische uitwisseling	Symmetrische uitwisseling
Multimedia	Multimedia (beperkt)	Gebaseerd op tekst

Gebaseerd op: Herring e.a. 2004, figuur 3.

De auteurs maken een onderscheid tussen twee typen blogs: dagboekachtige blogs en filter blogs.[46] Op het continuüm bevinden dagboekachtige blogs zich volgens de auteurs dichter bij standaard webpagina's, en bevinden filter blogs zich dichter bij asynchrone CMC websites. Mijns inziens kan het onderscheid dat de auteurs maken op het vlak van multimedia niet zo nadrukkelijk worden gemaakt. Op een groot aantal internetfora worden bijvoorbeeld afbeeldingen veelvuldig gepost.

Tot slot beschreef Blood de relatie tussen blogsoftware en blogs.[47] Veelal worden blogs gemaakt met blogsoftware. Vooral in de beginperiode werden blogs handmatig geprogrammeerd. De blogsoftware bepaalt op welke wijze blog posts binnen een blog kunnen worden weergegeven. Als voorbeelden van blogsoftwareapplicaties kunnen Blogger en Wordpress worden genoemd. Iedere applicatie beschikt over zijn eigen configuratiemogelijkheden.

Aldus kunnen blog posts worden beschouwd als een type webpagina dat over een vaste structuur en datum- en tijdstempel beschikt. Onderling kunnen blogs verschillen, doordat verschillende blogsoftware wordt gebruikt en opties verschillend zijn ingesteld.

[44] Een voorbeeld van asynchrone CMC websites zijn internetfora.
[45] Herring e.a. 2004, p. 10.
[46] Met filter blogs worden bedoeld de blogs die linken naar andere websites en deze websites becommentariëren. Zie voor een ander onderscheid Lomborg 2009.
[47] Blood 2004.

3.4 Gebruikers

De positie van de gebruikers van ontsluitingssystemen is onder meer bestudeerd door Wilson, Broder, Jansen e.a. en Kellar e.a..

Het onderzoek van Wilson richtte zich op het proces van het identificeren van de informatiebehoeften door gebruikers, het vervolgens zoeken naar deze informatie en tot slot het gebruiken of overnemen van deze informatie.[48] Al deze activiteiten samen noemt Wilson '*information behaviour*'. De auteur concludeert dat er naast zijn eigen model nauwelijks modellen bestaan om dit information behaviour in kaart te brengen. Wel kunnen in de literatuur specifiek op het vlak van het zoeken naar informatie ('*information seeking*') verschillende modellen worden onderkend. In het model van Wilson uit 1996 wordt ook met factoren van buiten de informatiewetenschap rekening gehouden.[49] Zo wordt in het model bijvoorbeeld met stress/'*coping*'[50] rekening gehouden, omdat deze factor ertoe kan leiden dat mensen niet naar informatie gaan zoeken. Een ander voorbeeld van een theorie waarmee rekening wordt gehouden is de '*risk/reward theory*'. Deze theorie kan verklaren waarom mensen een bepaalde informatiebron vaker gebruiken dan andere informatiebronnen. Tot slot wordt in een model van Wilson uit 1999 ook met de factor onzekerheid rekening gehouden.[51] Onzekerheid moet daarbij worden beschouwd vanuit de gebruiker. Wilson concludeert dat modellen niet als competitief, maar als complementair moeten worden beschouwd. Steeds moet worden afgevraagd of een model de werkelijkheid voldoende compleet weergeeft. In het onderhavige onderzoek wordt niet nader op voorgaande en soortgelijke modellen ingezoomd. Informatiebehoeften worden als gegeven beschouwd, en op basis van deze informatiebehoeften worden zoekvragen opgesteld.

Ook het onderzoek van Broder had betrekking op de informatiebehoeften van gebruikers.[52] Meer specifiek stelt Broder, in afwijking van de eerdere literatuur, dat informatiebehoeften niet per definitie ten grondslag liggen aan het gebruik van een internetzoekmachine. Het gebruik van een internetzoekmachine kan namelijk ook voortkomen uit het vinden van een website of het uitvoeren van een via het web tot stand te brengen activiteit. Broder onderscheidt daarom zoekvragen in de volgende drie klassen:

- informationele zoekvragen: om informatie te vinden over een onderwerp
- navigationele zoekvragen: om een specifieke website te vinden
- transactionele zoekvragen: ter uitvoering van een via het web tot stand te brengen activiteit

[48] Wilson 1999.
[49] Wilson 1999, figuur 7.
[50] Het aankunnen van het zoeken naar een bevrediging van de informatiebehoefte.
[51] Wilson 1999, figuur 13.
[52] Broder 2002.

Aldus liggen naar de mening van Broder informatiebehoeften slechts ten grondslag aan informationele zoekvragen. Broder heeft op basis van vragenlijsten en logbestanden onderzocht in welke verhouding de drie klassen zoekvragen voorkomen. Uit het onderzoek blijkt dat 40 à 50 procent van de zoekvragen informationeel is, dat 30 à 35 procent van de zoekvragen transactioneel is en dat 20 à 25 procent van de zoekvragen navigationeel is.
Jansen e.a. verrichtten in 2008 een vergelijkbaar onderzoek.[53] In hun onderzoek lag het percentage van informationele zoekvragen echter rond de 80% en lagen de percentages van navigationele en transactionele zoekvragen beide rond de 10%. Zoekvragen werden in het onderzoek van Jansen e.a. automatisch geclassificeerd. Uit handmatige controle van een steekproef bleek dat 26% van de 400 zoekvragen fout was geclassificeerd. Aannemende dat handmatige classificatie altijd juist is[54], zou het percentage informationele zoekvragen rond de 65% bedragen, het percentage navigationele zoekvragen rond de 15% bedragen en het percentage transactionele zoekvragen rond de 20% bedragen. Navigationele en transactionele zoekvragen maken geen onderdeel uit van het onderhavige onderzoek. Uitsluitend informationele zoekvragen zullen worden opgesteld.

Het onderzoek van Kellar e.a. richtte zich niet op de informatiebehoeften van gebruikers, maar op hun informatiezoekgedrag.[55] De onderzoekers onderzochten welke browserfunctionaliteiten tijdens webgebaseerde taken op het terrein van het opzoeken van informatie worden gebruikt. Daarnaast onderzochten de onderzoekers of het gebruik van browserfunctionaliteiten per taak (vinden van feiten, informatie verzamelen, rondkijken, transacties verrichten) verschilde. De onderzoekers onderzochten de vragen door de browserinteracties van proefpersonen te loggen, en daarnaast de proefpersonen te vragen een elektronisch dagboek bij te houden om hun gedrag te beschrijven en te classificeren. In totaal deden 21 studenten mee aan het onderzoek dat een week duurde. 178 browserfuncties werden geregistreerd. Uit het onderzoek bleek dat bij de taak van het informatie verzamelen het meeste van de browserfuncties gebruik werd gemaakt. Zo werden vaker pagina's gebookmarked voor later en werd de Google Toolbar vaker gebruikt. Verder waren de proefpersonen ten opzichte van de andere taken bij het informatie verzamelen langer bezig en bezochten zij meer pagina's. Ten opzichte van de taak van het vinden van feiten waren de zoekvragen bij de taak van het informatie verzamelen aanmerkelijk korter. In het onderhavige onderzoek wordt het informatiezoekgedrag niet bestudeerd aan de hand van proefpersonen. Zoals reeds opgemerkt zullen informatiebehoeften als gegeven worden aangenomen en worden de zoekvragen door de onderzoeker zelf opgesteld en uitgevoerd. Het gebruik van browserfunctionaliteiten wordt niet nader beschouwd.

Onderzoek naar de vraag wie de internetters die blogs zoeken en lezen zijn, kan nauwelijks worden gevonden in de wetenschappelijke literatuur. Wel kan op basis van het onderzoek van Boyd worden

[53] Jansen e.a. 2008.
[54] Hiervan uitgaande zouden achter dezelfde zoekvraag bij verschillende mensen niet verschillende bedoelingen schuil kunnen gaan.
[55] Kellar e.a. 2007.

geconcludeerd dat een aantal van de bezoekers van blogs ook zelf een blog bijhoudt.[56] Ten aanzien van de mate waarin bloggers zich onder de lezers en zoekers van blogs bevinden kan echter geen onderzoek worden gevonden in de literatuur. In hoofdstuk 4 wordt nader ingegaan op onderzoek naar de zoekvragen die gebruikers in blogzoekmachines uitvoeren.

Aldus kan worden geconcludeerd dat de gebruikers van ontsluitingssystemen in het algemeen, en van blogzoekmachines in het bijzonder, van belang zijn bij de werking van het systeem. In hoeverre de gebruikers van blogzoekmachines verschillen van internetters in het algemeen kan niet worden gevonden in de wetenschappelijke literatuur.

3.5 Systeem

Een zoekmachine is een ontsluitingssysteem dat representaties van documenten bevat. Op basis van de zoekvragen van de gebruikers presenteert het systeem de meest relevante resultaten aan de gebruiker. Aangezien het internet uit oneindig veel pagina's (en andere objecten) bestaat, bevatten internetzoekmachines slechts representaties van een deel van de op het internet aanwezige pagina's. Bovendien is het mogelijk dat representaties verouderd zijn. Een algemene internetzoekmachine ontdekt meestal nieuwe pagina's door verwijzingen (links) vanaf de reeds bekende pagina's naar de nog niet bekende pagina's te volgen. Een bijzondere vorm hiervan is het volgen van links in RSS-feeds. Een andere manier waarop internetzoekmachines achter het bestaan van nieuwe pagina's kan komen is attendering (veelal door middel van 'pinging'). Vooral blogzoekmachines, de zoekmachines die zich vrijwel uitsluitend op blogs en/of nieuws richten, bieden tegenwoordig mogelijkheden hiertoe.

Lewandowski en Höchstötter hebben een kader opgesteld waarbinnen de kwaliteit van internetzoekmachines op vier verschillende gebieden kan worden uitgedrukt.[57] Deze vier deelgebieden zijn:

- **Indexkwaliteit**. Met dit deelgebied drukken de auteurs het belang van het vinden van relevante, uitgebreide verzameling resultaten uit. Binnen dit deelgebied onderscheiden de auteurs de volgende maatstaven:
    - '*Web coverage*' (dekking in het algemeen)
    - '*country bias*' (worden pagina's uit een land bevoordeeld of benadeeld bij de opname in de zoekmachine?)
    - recentheid van gerepresenteerde documenten

[56] Boyd 2006. Zie ook Blood 2004.
[57] Lewandowski en Höchstötter 2008, p. 319-320.

- **Kwaliteit van de resultaten**. Binnen dit deelgebied moeten de klassieke retrieval tests worden geplaatst. Maatstaven zoals precisie en vangst zijn gebruikelijk, maar de auteurs geven aan dat ook andere factoren van belang kunnen zijn. Daarbij geven de auteurs als voorbeeld: de mate waarin een zoekmachine in verhouding tot andere zoekmachines unieke resultaten biedt.
- **Kwaliteit van de zoekmogelijkheden**. Onder dit deelgebied verstaan de auteurs de beschikbaarheid van uitgebreide mogelijkheden om de wijze waarop wordt gezocht in te stellen (bijvoorbeeld de aanwezigheid van een scherm voor geavanceerd zoeken). Ook de beschikbaarheid van verfijnde mogelijkheden om de zoekvraag op te stellen valt onder dit deelgebied.
- **Zoekmachinegebruiksvriendelijkheid** ('*search engine usability*'). Binnen dit deelgebied kan de kwaliteit worden gemeten aan de hand van feedback van het gedrag van de gebruikers. Deze feedback kan worden verkregen aan de hand van gebruikersonderzoek, laboratoriumtests en transactieloganalyse. Binnen dit deelgebied valt bijvoorbeeld het onderzoek naar de vraag hoe effectief en efficiënt gebruikers zoekmachines gebruiken.

In verband met de beperkte tijd richt het onderhavige onderzoek zich uitsluitend op de kwaliteit van de resultaten van zoekmachines: in hoeverre bieden zoekmachines relevante resultaten? Onderzoek op dit punt is onder meer verricht door Tague-Sutcliffe. Tague-Sutcliffe bestudeerde uitvoerig de wijze waarop information retrieval experimenten zijns inziens het beste kunnen plaats vinden.[58] In zijn artikel presenteerde de auteur een stappenplan hiervoor. Tevens analyseerde de auteur de maatstaven om de kwaliteit van de resultaten te meten. Gegrond op het artikel van Tague-Sutcliffe wordt in paragraaf 6.6 de keuze van de maatstaven in het experiment toegelicht.

De kwaliteit van de resultaten van internetzoekmachines, tevens te omschrijven als de effectiviteit van internetzoekmachines, is door een groot aantal onderzoekers bestudeerd. Lewandowski en Morisson hebben afzonderlijk de experimenten in de wetenschappelijke literatuur vergeleken.[59] Opgemerkt moet worden dat bij onderzoek naar de kwaliteit van de resultaten regelmatig proefpersonen worden gebruikt. Recent onderzochten Bar-Ilan e.a. bijvoorbeeld met behulp van 65 proefpersonen de ordening van drie grote internetzoekmachines.[60] Het onderhavige onderzoek wordt in verband met praktische problemen en de beperkte tijd in zijn geheel door de onderzoeker uitgevoerd.

Onderzoek naar de mate waarin lezers en zoekers van blogs blogzoekmachines gebruiken om blogs te vinden kan niet worden gevonden in de wetenschappelijke literatuur. Uit reacties op verschillende blogs kan wel worden afgeleid dat sommige mensen vrijwel nooit blogzoekmachines gebruiken. Deze

[58] Tague-Sutcliffe 1992.
[59] Lewandowski 2008, tabel 1 en Morisson 2008, tabel 1.
[60] Bar-Ilan e.a. 2009.

internetters geven aan algemene zoekmachines zoals Google Search of nieuwere, specifieke zoekmachines zoals Twitter Search en FriendFeed Search te gebruiken.[61]

Tevens kan in de wetenschappelijke literatuur nauwelijks evaluerend onderzoek naar de kwaliteit van de resultaten van blogzoekmachines worden gevonden. Het enige onderzoek waarin de kwaliteit van resultaten (mede) aan bod komt is het onderzoek van Thelwall en Hasler.[62] Dit explorerend onderzoek is echter te beperkt om conclusies uit af te leiden. Aldus kan worden geconcludeerd dat evaluerend onderzoek naar de kwaliteit van de resultaten van blogzoekmachines zeer welkom is.

3.6 Conclusie

Drie elementen spelen binnen een internetzoekmachine en ieder ander ontsluitingssysteem een belangrijke rol: informatie, de gebruikers en een systeem dat de informatiebehoeften van gebruikers en de beschikbare informatie matcht. In tegenstelling tot de standaard webpagina, kent de blog een vaste structuur waarin de blog posts in omgekeerd chronologische volgorde geordend zijn. In hoeverre de gebruikers van blogzoekmachines verschillen van internetters in het algemeen is niet bekend. De effectiviteit van internetzoekmachines in het algemeen is reeds uitgebreid onderzocht. Ten aanzien van de effectiviteit van blogzoekmachines kan in de wetenschappelijke literatuur echter nauwelijks evaluerend onderzoek worden gevonden. Alvorens aan het experiment toe te komen in hoofdstuk 6 en 7, moeten in hoofdstuk 4 en 5 nog enkele zaken worden verdiept. In hoofdstuk 4 wordt allereerst het informatiezoekgedrag van gebruikers van blogzoekmachines onderzocht: ‘Waarnaar en hoe zoeken internetters met blogzoekmachines?’ In hoofdstuk 5 wordt vooral het systeem van de zoekmachine belicht: ‘Wat zijn de verschillen en overeenkomsten tussen de functionaliteiten van blogzoekmachines onderling, en ten opzichte van Google Search?’.

[61] Zie bijvoorbeeld Rubel 2008, Rosenberg 2008, Diana 2008 en A. Agarwal 2009.
[62] Thelwall & Hasler 2007.

4 INFORMATIEZOEKGEDRAG

4.1 Inleiding

In het vorige hoofdstuk werd geconcludeerd dat de omgang van internetters met blogzoekmachines nog nader moet worden onderzocht. Het informatiezoekgedrag van internetters wordt daarom in dit hoofdstuk aan de hand van de volgende vraag bestudeerd: 'Waarnaar en hoe zoeken internetters met blogzoekmachines?' Alvorens nader in te gaan op de opbouw van dit hoofdstuk, moet worden verduidelijkt dat in het spraakgebruik onder 'het zoeken naar blogs' meestal niet 'het zoeken naar blogs' wordt verstaan, maar vaker het 'zoeken naar blog posts'.[63] In het experiment van hoofdstuk 6 en 7 staat puur het zoeken naar blog posts centraal.
In paragraaf 4.2 wordt onderzocht waarnaar internetters met blogzoekmachines zoeken. Vervolgens wordt in paragraaf 4.3 onderzocht hoe internetters met blogzoekmachines zoeken. Paragraaf 4.4 sluit het hoofdstuk af met een conclusie.

4.2 Zoekvragen in blogzoekmachines

In de wetenschappelijke literatuur kunnen twee verschillende onderzoeken worden gevonden met betrekking tot de zoekvragen die internetters in blogzoekmachines uitvoerden. Het eerste onderzoek betreft het onderzoek van Mishne en De Rijke naar de (inmiddels inactieve) zoekmachine Blogdigger.[64] Dit onderzoek vond plaats aan de hand van de logbestanden van Blogdigger uit de maand mei 2005. Deze logbestanden geven inzicht in de zoekvragen (de 'queries') die in deze maand aan de zoekmachine zijn verzonden.
Mishne en De Rijke zijn van mening dat het in paragraaf 3.4 door Broder gemaakte onderscheid van zoekvragen in blogzoekmachines minder geschikt is. De onderzoekers merken op dat transactionele zoekvragen geen natuurlijke categorie zoekvragen voor een blogzoekmachine zijn. Tevens merken ze op dat een internetter voor navigationele zoekvragen eerder een algemene zoekmachine zal gebruiken dan een blogzoekmachine. Deze gedachten brachten de onderzoekers tot de hypothese dat gebruikers van blogzoekmachines voor het merendeel informationele zoekvragen in blogzoekmachines uitvoeren. Een scan van de logbestanden van Blogdigger bevestigde deze hypothese.

Mishne en De Rijke analyseerden een willekeurige set van 1000 zoekvragen, waarvan de ene helft ad-hoc zoekvragen betrof en de andere helft filtering zoekvragen betrof. Ad-hoc zoekvragen zijn de zoekvragen die gebruikers rechtstreeks via de website van de blogzoekmachine uitvoeren. Filtering zoekvragen zijn de zoekvragen die buiten de interface van de blogzoekmachine om en in de regel

[63] Seo en Croft 2008, 1053.
[64] Mishne 2007, p. 183, en Mishne & De Rijke 2006.

periodiek worden verstuurd, en die vervolgens in bijvoorbeeld XML of JSON bij de gebruiker terechtkomen.
In de set van de 500 ad-hoc zoekvragen bleek 52% van de zoekvragen te bestaan uit namendragende entiteiten, zoals namen van mensen, producten of bedrijven. In de set van filtering zoekvragen bleek 78% van de zoekvragen te bestaan uit namendragende entiteiten. Meer dan 80% van de zoekvragen in beide sets bleek te bestaan uit ofwel zeer bekende namen (bijvoorbeeld 'Microsoft' of 'Bush'), ofwel zeer onbekende namen van individuen en bedrijven. Een andere populaire categorie zoekvragen betrof locatienamen (veelal Amerikaanse steden). De zoekvragen die niet uit namendragende entiteiten bestonden bleken grotendeels (25% van de ad-hoc zoekvragen en 18% van de filtering zoekvragen) te bestaan uit hogere concepten en onderwerpen. Als voorbeelden van de ad-hoc zoekvragen uit de logbestanden kunnen worden genoemd: 'stock trading', 'linguists', 'humor', 'gay rights' en 'islam'. De filtering zoekvragen in deze categorie waren veelal technologiegerelateerd. De rest van de zoekvragen bestond uit adult-georiënteerde zoekvragen (vrijwel uitsluitend ad-hoc zoekvragen), URL-zoekvragen, en diverse zoekvragen zonder specifieke kenmerken.

Naast deze willekeurige set van zoekvragen hebben Mishne en De Rijke tevens een set met slechts unieke zoekvragen nader onderzocht. De set bestond uit de 400 meest voorkomende zoekvragen binnen de set unieke zoekvragen. Opnieuw bestond de helft uit ad-hoc zoekvragen en de helft uit filtering zoekvragen. De verdeling tussen de categorieën zoekvragen bleek licht te verschillen. Bij deze verzameling unieke zoekvragen bleek 45% van de ad-hoc zoekvragen en 66% van de filtering zoekvragen uit namendragende entiteiten te bestaan. Concepten en technologieën bleken 30% van de ad-hoc zoekvragen en 28% van de filtering zoekvragen te vormen. Adult-georiënteerde ad-hoc zoekvragen bleken in deze set beduidend meer voor te komen dan in de willekeurige set.

Het tweede onderzoek betreft het onderzoek van Sun e.a..[65] De onderzoekers hebben gedurende de periode november 2006 – november 2007 iedere drie uur de top 15 van populairste zoekvragen bij Technorati opgevraagd.[66] In totaal hebben de onderzoekers 44594 zoekvragen verzameld. Na het verwijderen van leestekens bleven 1530 verschillende zoekvragen over. 63% van deze zoekvragen kwam minder dan 10 keer voor. Een klein aantal zoekvragen was gedurende een lange periode populair. De zoekvraag 'youtube' kwam bijvoorbeeld 2766 keer voor, en was tijdens 346 dagen van de onderzochte 388 dagen populair. De top 20 van meest populaire zoekvragen delen de onderzoekers grofweg in drie categorieën in: websites, personen en onderwerpen/concepten. Bijna de helft van de zoekvragen bestond uit namen van websites (zoals 'YouTube', 'MySpace' en 'Google'). Een kleiner deel van de top 20 van meest populaire zoekvragen bestond uit namen van personen (zoals 'Ron Paul', 'Paris Hilton' en 'Noelia'). Een nog kleiner deel van de top 20 van meest populaire zoekvragen

[65] Sun e.a. 2008.
[66] <http://technorati.com/pop/>

bestond uit onderwerpen/concepten (zoals ‘web 2.0’, ‘video’ en ‘music’). Buiten de drie categorieën kwam tot slot ook de term ‘iphone’ in de top 20 van meest populaire zoekvragen voor.

Vergeleken met het onderzoek van Mishne en De Rijke concludeerden Sun e.a. dat uit beide onderzoeken de conclusie kan worden getrokken dat de gebruikers zeer vaak namendragende entiteiten als zoekvraag kozen, en dat deze vaak aan gebeurtenissen gerelateerd waren. Als voorbeelden hiervan in hun eigen onderzoek noemden Sun e.a. de zoekvraag ‘harry potter’ (vanwege de release van het boek), ‘saddam hussein’ (vooral vanwege de executie video) en ‘Windows Vista’ (vanwege de release van het softwareproduct). Als verschil met het onderzoek van Mishne en De Rijke noemden Sun e.a. het feit dat namen van websites in grotere aantallen aanwezig waren in hun eigen onderzoek. Mijns inziens kan dit verschil gemakkelijk worden verklaard uit de lengte van websitenamen. Populaire websitenamen zijn korter, waardoor ze vaker in de top 15 van populaire zoekvragen kunnen voorkomen. Mishne en De Rijke onderzochten niet alleen de populaire zoekvragen, maar ook de zoekvragen die minder vaak werden verstuurd.

In aanvulling op voorgaande twee onderzoeken uit de wetenschappelijke literatuur, kan het marktonderzoek van Universal McCann worden geraadpleegd.[67] De onderzoekers van Universal McCann ondervroegen een afspiegeling van de wereldwijde, actieve internetbevolking over de onderwerpen blogs die men las. Uit het onderzoek kan worden afgeleid dat 63,5% van de blogs lezende mensen aangaf persoonlijke, dagboekachtige blogs te lezen. 38,9% van de blogs lezende mensen gaf aan ‘*friend blogs*’ te lezen.[68] De echt onderwerpsgerelateerde categorieën scoorden lager: het onderwerp muziek scoorde 33,1%, terwijl het onderwerp nieuws / actuele zaken 29,1% scoorde. De onderwerpen meningen over producten/merken, film/tv, computers, reizen en technologie scoorden rond de 25%.

Aldus kan worden geconcludeerd dat internetters blogzoekmachines vooral voor informationele zoekvragen lijken te gebruiken. Daarbij zijn internetters die deze zoekmachines gebruiken vooral op zoek naar informatie over de zeer bekende personen en organisaties. Onder de zoekvragen die niet uit namendragende entiteiten bestonden, konden vooral hogere concepten en onderwerpen worden gevonden. Blogs over de onderwerpen muziek, nieuws/actuele zaken, meningen over producten/merken, film/tv, computers, reizen en technologie lijken internetters het vaakst te lezen.

[67] Universal McCann 2008, p. 21.
[68] Het verschil tussen vorengenoemde categorieën werd niet toegelicht.

4.3 Wijze van zoeken met blogzoekmachines

Naast het onderzoeken van de zoekvragen die internetters in blogzoekmachines uitvoeren, kan ook worden onderzocht hoe deze internetters met blogzoekmachines zoeken. Op dit punt kan echter nauwelijks onderzoek worden gevonden in de wetenschappelijke literatuur. Alleen Mishne en De Rijke besteden zijdelings aandacht aan de lengte van de zoekvragen die internetters in Blogdigger uitvoerden. De onderzoekers stelden vast dat een ad-hoc zoekvraag gemiddeld uit 2,44 woorden bestond.[69] Hiermee concludeerden de onderzoekers dat dit gemiddelde overeenstemt met de gemiddelden uit onderzoek naar algemene internetzoekmachines.

Ten aanzien van de wijze waarop internetters algemene internetzoekmachines gebruiken, kan wel onderzoek worden gevonden. Allereerst kan op het punt van de lengte van zoekvragen het onderzoek van marktonderzoeksbureau Hitwise worden genoemd. Hitwise zag in de periode januari 2008 – januari 2009 een lichte toename van het aantal woorden dat mensen gebruikten voor hun zoekvragen in algemene internetzoekmachines.[70] In januari 2009 bestond 20% van de zoekvragen uit 1 woord, 24% van de zoekvragen uit 2 woorden, 22% van de zoekvragen uit 3 woorden, 15% van de zoekvragen uit 4 woorden en de rest (19%) uit 5 of meer woorden.

Naast de lengte van zoekvragen is tevens het aantal resultatenpagina's dat gebruikers van algemene internetzoekmachines bekijken onderzocht. Jansen e.a. concludeerden in 2000 dat gebruikers in de zoekmachine Excite gemiddeld 2,35 resultatenpagina's van 10 resultaten opvroegen.[71] 58% van de gebruikers keek echter niet verder dan de eerste pagina. Recenter onderzoek op dit punt kan niet worden gevonden. Mijns inziens zal het aantal gebruikers dat niet verder kijkt dan de top 10 resultaten tegenwoordig nog hoger liggen, omdat de kwaliteit van de resultaten van zoekmachines sindsdien is verbeterd, en daarnaast de zoekmachine sindsdien veel meer in het gebruik van de computer is ingebed en mensen de zoekmachine meer voor eenvoudigere vragen zijn gaan gebruiken.

Tot slot is ook onderzoek gedaan naar het aantal zoekvragen per sessie. Jansen e.a. concludeerden in 2000 dat 67% van de gebruikers één zoekvraag in de algemene internetzoekmachine Excite uitvoerden, en dat 19% van de gebruikers twee zoekvragen uitvoerden. De rest van de gebruikers (rond de 14%) voerde meer dan twee zoekvragen uit. Recenter onderzoek op dit punt kan niet worden gevonden. Mijns inziens zal het aantal gebruikers dat per sessie slechts één zoekvraag uitvoert tegenwoordig nog hoger liggen, omdat de kwaliteit van de resultaten van zoekmachines sindsdien is verbeterd, en daarnaast de zoekmachine sindsdien veel meer in het gebruik van de computer is ingebed en mensen de zoekmachine meer voor eenvoudigere vragen zijn gaan gebruiken. Desondanks zullen

[69] Indien alleen de unieke ad-hoc zoekvragen werden geanalyseerd bedroeg het gemiddelde 2,71 woorden.
[70] Vgl. Hitwise 2009.
[71] Jansen e.a. 2000, p. 214.

veel internetters in sommige gevallen nog steeds uitgebreid zoekvragen binnen een sessie verfijnen en verbreden.

Aldus kan worden geconcludeerd dat ook ten aanzien van de wijze waarop internetters met blogzoekmachines zoeken nauwelijks wetenschappelijk onderzoek bestaat. Geconcludeerd kan worden dat internetters in blogzoekmachines gemiddeld in 2,44 woorden hun zoekvraag opstelden. Ook kan uit onderzoek uit 2000 worden geconcludeerd dat gebruikers van algemene internetzoekmachines in het merendeel van de uitgevoerde zoekvragen slechts de eerste pagina opvroegen. Tot slot kan uit hetzelfde onderzoek worden geconcludeerd dat tweederde van de internetters in algemene internetzoekmachines slechts één zoekvraag uitvoert.

4.4 Conclusie

Op het gebied van de vraag waarnaar en hoe internetters met blogzoekmachines zoeken kan in de literatuur nauwelijks wetenschappelijk onderzoek worden gevonden. Internetters lijken blogzoekmachines vooral voor informationele zoekvragen te gebruiken. Daarbij zijn internetters die deze zoekmachines gebruiken vooral op zoek naar informatie over de zeer bekende personen en organisaties, en in iets minder mate hogere concepten en onderwerpen. Blogs over de onderwerpen muziek, nieuws / actuele zaken, meningen over producten/merken, film/tv, computers, reizen en technologie lijken internetters het vaakst te lezen.
Ook ten aanzien van de wijze waarop internetters met blogzoekmachines zoeken bestaat nauwelijks wetenschappelijk onderzoek. Geconcludeerd kan worden dat internetters in blogzoekmachines gemiddeld in 2,44 woorden hun zoekvraag opstelden. Ook kan uit onderzoek uit 2000 worden geconcludeerd dat gebruikers van algemene internetzoekmachines in het merendeel van de uitgevoerde zoekvragen slechts de eerste pagina opvroegen. Tot slot kan uit hetzelfde onderzoek worden geconcludeerd dat tweederde van de internetters in algemene internetzoekmachines slechts één zoekvraag uitvoerde.
Mede gebaseerd op deze uitkomsten zijn de zoekvragen voor het experiment van hoofdstuk 6 en 7 opgesteld. Aldus richten de zoekvragen zich op de zeer bekende personen, organisaties, producten, locaties en hogere concepten en onderwerpen. Bij het opstellen wordt geprobeerd de zoekvragen op een lengte van ongeveer 2,5 woorden te houden. Per zoekvraag worden slechts de top 10 resultaten op relevantie beoordeeld. Tot slot wordt bij iedere omschreven informatiebehoefte slechts één zoekvraag opgesteld.

5 FUNCTIONALITEITEN BLOGZOEKMACHINES

5.1 Inleiding

Om de effectiviteit van blogzoekmachines zorgvuldig te kunnen evalueren, zal moeten worden onderzocht wat de belangrijkste functionaliteiten van de in het experiment te betrekken blogzoekmachines zijn. Omdat de blogzoekmachines in het experiment van hoofdstuk 6 en 7 tevens met Google Search worden vergeleken, zullen ook in dit hoofdstuk de functionaliteiten van Google Search worden onderzocht. Het onderzoek in dit hoofdstuk vindt plaats aan de hand van de volgende deelvraag: 'Wat zijn de verschillen en overeenkomsten tussen de functionaliteiten van blogzoekmachines onderling, en ten opzichte van Google Search?'. Om de verschillen in effectiviteit zorgvuldig te kunnen analyseren, moet hierbij tevens de achtergrond waartegen de functionaliteiten van de verschillende zoekmachines zijn ontstaan, worden belicht.
In het experiment van hoofdstuk 6 en 7 worden vier blogzoekmachines betrokken: Google Blog Search, Technorati, Icerocket Blog Search en Twingly Blog Search. Tevens wordt de algemene internetzoekmachine Google Search betrokken. Daarom zal dit hoofdstuk zich specifiek op deze vijf zoekmachines richten. Allereerst wordt in paragraaf 5.2 beschreven hoe de blogzoekmachines en Google Search zijn ontstaan, en wie de organisaties achter deze zoekmachines zijn. In paragraaf 5.3 wordt een overzicht van de belangrijkste overeenkomsten en verschillen tussen de blogzoekmachines gegeven. Paragraaf 5.4 sluit het hoofdstuk af met een conclusie.

5.2 Ontstaan en ontwikkelaars

Hieronder wordt het ontstaan van de zoekmachines Google Search, Google Blog Search, Technorati, Icerocket en Twingly kort toegelicht. Daarbij worden tevens de organisaties achter deze zoekmachines belicht.

Google Search en Google Blog Search

Google Search, de algemene zoekmachine van Google, is in 1998 opgericht en in de meeste landen in de wereld uitgegroeid tot de meest gebruikte zoekmachine. Google omschrijft zijn missie als volgt: '*to organize the world's information and make it universally accessible and useful*'[72]. Sinds de start met de algemene zoekmachine in 1998 heeft de onderneming allerlei producten gelanceerd, waaronder een nieuwszoekmachine, een emaildienst (Gmail) en een browser (Chrome). In 2008 maakte de onderneming voornamelijk via advertentie-inkomsten een winst van 6,6 miljard dollar.
De blogzoekmachine Google Blog Search is in 2005 opgericht door Google. Sinds 2007 integreert Google de content van de blogzoekmachine en andere specifieke zoekmachines (Afbeeldingen,

[72] <http://www.google.com/intl/en/corporate/>

Nieuws etc.) in de algemene zoekmachine.[73] Hiermee biedt Google de mogelijkheid van Universal Search: alle content kan in één keer worden doorzocht.

Technorati

Technorati startte in 2002 met een dienst om de inkomende links naar blogs bij te kunnen houden. Sinds juni 2003 biedt het bedrijf een blogzoekmachine aan. Technorati richt zich echter niet puur op het vinden van blog posts. Het bedrijf omschrijft zijn activiteiten namelijk zelf als het verzamelen, organiseren en distribueren van de '*global online conversation*'.[74]

Icerocket

Icerocket startte in 2004 met een metazoekmachine. Sinds het najaar van 2004 biedt Icerocket tevens een blogzoekmachine. Deze blogzoekmachine wordt sinds halverwege 2005 als de standaard zoekmachine gepresenteerd. Naast de blogzoekmachine en de metazoekmachine (die zich tegenwoordig slechts op de resultaten van Google Search lijkt te baseren), beschikt Icerocket over nog enkele zoekmachines (Twitter, MySpace, News en Pictures). Deze laatste zoekmachines lijken zich echter te baseren op databases van externe zoekmachines.

Twingly Blog Search

Twingly Blog Search[75] is de jongste van de vier blogzoekmachines. Deze Zweedse zoekmachine bestaat sinds begin 2008. Vanaf het begin biedt Twingly niet alleen een zoekmachine voor het publiek, maar ook automatisch gegenereerde overzichten van relevante blog posts die krantenuitgevers en andere nieuwsleveranciers bij artikelen op hun website kunnen plaatsen. Hierdoor kunnen de lezers van deze artikelen over het betreffende onderwerp verder lezen op blogs.[76] In het begin richtte de blogzoekmachine zich vooral op Europese blogs, maar tegenwoordig biedt de blogzoekmachine een ruime dekking van blogs uit de hele wereld. Naast de blogzoekmachine heeft Twingly in december 2008 een Microblog Search[77] zoekmachine gelanceerd, welke microblogs (zoals Twitter en Jaiku) doorzoekt.[78]

---

[73] Vgl. Mayer 2007.
[74] Vgl. <http://technoratimedia.com/>
[75] <http://www.twingly.com/>
[76] Vgl. <http://www.twingly.com/partners>
[77] <http://www.twingly.com/microblogsearch>
[78] In dit onderzoek wordt Twingly Microblog Search buiten beschouwing gelaten, omdat de posts op microblogs mijns inziens veelal te beknopt zijn om aan de informatiebehoeften van de zoekende internetter te kunnen voldoen. Daarnaast leidt de beknoptheid van microblog posts mijns inziens veelal tot een tekort aan context om zoekmachines in staat te stellen aan de internetter de meest relevante posts te tonen. Tot slot bemoeilijkt de verkorte weergave van links (door middel van diensten zoals Tinyurl) binnen deze microblog posts het beoordelen door de internetter of via deze links lezenswaardig materiaal kan worden gevonden.

5.4 Verschillen en overeenkomsten in functionaliteiten

De belangrijkste overeenkomsten en verschillen tussen de te onderzoeken blogzoekmachines en Google Search zijn onderzocht door de interfaces te bestuderen en enkele zoekacties uit te voeren. De overeenkomsten en verschillen zijn weergegeven in Tabel 2.

Tabel 2 – Overzicht verschillen en overeenkomsten tussen blogzoekmachines

Zoekmachines Functies/opties	Google Blog Search	Technorati	Icerocket Blog Search	Twingly Blog Search	Google Search
Terrein	Alle	Alle	Alle	Alle	Alle
Standaard taal waarin gezocht wordt	Alle	Alle[79]	Alle	Alle	Alle
Standaard sortering	Relevantie	Datum[80]	Datum	Relevantie	Relevantie
Overige sorteeropties	Datum	-	-	Datum Inlinks Inlinks (website) Recommendations Recommendations (website)	Datum[81]
Sorteermethoden	Aflopend	Aflopend	Aflopend	Aflopend Oplopend	Aflopend
Filtering mogelijk op basis van:	Periode Bron Taal Auteur Link[82]	Taal[83] Link[84] URL Tag Onderwerp van blog Autoriteitsgraad[85]	Periode Bron Taal Auteur Link Tag	Periode[86] Bron Taal Link URL Spam	Taal Link URL Bestandstype Gebruiksrechten Regio
Maximum aantal weer te geven resultaten	Ongeveer 1000	Onbekend[87]	1000	1000	Ongeveer 1000
Getal met punt of komma in zoekterm mogelijk	Ja	Ja	Ja	Nee, wordt genegeerd[88]	Ja
Reacties op blog posts doorzoekbaar?	Nee	Nee	Nee	Nee	Ja
Feeds of volledige blog posts doorzoekbaar?	Volledige blog posts[89]	Feeds[90]	Volledige blog posts[91]	Feeds[92]	Volledige webpagina's
Trendgrafiek	Nee	Ja	Ja	Nee	Ja
RSS-feeds van zoekvragen	Ja	Ja	Ja	Ja	Deels[93]

[79] Sinds begin 2009.
[80] Al toont Technorati vooraf soms eerst een top 10 van blog posts die aanmerkelijk ouder zijn. Aldus ondervonden door te zoeken op 'blog' (25 mei 2009).
[81] Sinds mei 2009. Een eerste blik laat zien dat een groot deel van de webpagina's niet wordt weergegeven. Daarnaast wordt de datum niet altijd juist vastgesteld.
[82] Resultaten zijn sinds oktober 2008 matig. Volgens Google heeft een update dit probleem sterk verminderd. Toch kan nog steeds (6 juni 2009) worden geconstateerd dat soms irrelevante resultaten worden getoond. Vgl. Hylton 2009.
[83] In de geavanceerd zoeken interface kan echter geen taal worden geselecteerd.
[84] Het uitvoeren van de zoekvraag via geavanceerd zoeken werkt momenteel niet.
[85] In de geavanceerd zoeken interface kun je echter geen autoriteitsgraad selecteren.
[86] Tot een maand terug.
[87] N.B. Uit vooronderzoek blijkt dat het Technorati niet altijd lukt meer dan de top 10 resultaten te tonen. Indien resultatenpagina's rechtstreeks worden aangeroepen wordt ongeacht het nummer van de resultatenpagina slechts de top 10 getoond.
[88] Vgl. <http://www.twingly.com/search?q=%22Firefox+3.0.6+browser%22>
[89] Vgl. bijvoorbeeld de resultaten op basis van zoektermen uit artikelen op <http://www.velonews.com> (site met onvolledige feed).
[90] Vgl. bijvoorbeeld de resultaten op basis van zoektermen uit artikelen op <http://www.velonews.com> (site met onvolledige feed).
[91] Vgl. bijvoorbeeld de resultaten op basis van zoektermen uit blog posts op <http://passievoorhanden.web-log.nl> (blog met onvolledige feed).
[92] Vgl. bijvoorbeeld de resultaten op basis van zoektermen uit blog posts op <http://www.hetzuur.nl> (blog met onvolledige feed).
[93] Vgl. <http://www.google.com/alerts> en <http://www.googlealert.com>

5.4.1 Opmerkingen

Aanvullend op tabel 2 dient het functioneren van individuele zoekmachines nog te worden toegelicht.

Google Blog Search

Sinds het laatste kwartaal van 2008 kent Google Blog Search problemen met het indexeren van blog posts, waardoor ook de content van 'blog rolls' (linkoverzichten naar andere blogs) geregeld tot de inhoud van de blog post wordt gerekend.[94] Gevolg is dat de blogzoekmachine niet relevante blog posts als resultaten kan tonen. Jeremy Hylton van het Google Blog Search Team liet eind december 2008 in een post op het forum van Google Blog Search weten dat Google aan dit probleem werkt.[95] Op 27 maart 2009 gaf Hylton in dezelfde thread aan dat de ranking is aangepast, waardoor resultaten minder vaak op basis van de content van blog rolls als relevant worden beschouwd. Tot slot liet Hylton op 13 april 2009 weten dat de zoekmachine is aangepast ten aanzien van blog rolls bij gebruik van het 'link:' argument.[96] Voor het experiment van hoofdstuk 6 en 7 moet worden benadrukt dat Google Blog Search is geraadpleegd op de dagen 8 tot en met 13 maart 2009. Aldus is het mogelijk dat de kwaliteit van de resultaten sindsdien aanmerkelijk is verbeterd of verslechterd.[97]

Google Search

Google Search ordent de resultaten standaard op PageRank. Google omschrijft deze maatstaf zelf als volgt:

> "*PageRank evaluates all of the sites linking to a web page and assigns them a value, based in part on the sites linking to them. By analyzing the full structure of the web, Google is able to determine which sites have been "voted" the best sources of information by those most interested in the information they offer.*"[98]

De PageRank is dus een cijfer dat tegelijkertijd het aantal links naar een pagina en het belang van deze links naar een pagina uitdrukt.

Google Search kent geen goed werkende mogelijkheden om een zoekvraag naar periode waarin gezocht wordt te beperken, omdat de datum waarop een webpagina is toegevoegd of is gewijzigd veelal niet vrij beschikbaar is. Uit vooronderzoek is bekend dat Google Search geregeld aan oude webpagina's een recente datum toekent, terwijl deze webpagina's sindsdien niet gewijzigd zijn.

---

[94] Owens 2008, Rubel 2008, Rosenberg 2008, Diana 2008 en Gray 2009.
[95] Hylton 2008.
[96] Hylton 2009.
[97] De rangschikking op relevantie lijkt sindsdien minder waarde aan blog posts van meer dan zes tot twaalf maanden oud te hechten.
[98] <http://www.google.com/intl/en/corporate/tenthings.html>

Twingly Blog Search

In Twingly Blog Search moet worden gekozen tussen standaard zoeken en spam-free zoeken. Blogs worden goedgekeurd alvorens ze binnen spam-free Blog Search verschijnen. Deze verzameling van goedgekeurde blogs lijkt tot stand te komen op basis van links in blog posts van reeds goedgekeurde blogs. In de FAQ op de website van Twingly wordt de wijze waarop met goedgekeurde blogs wordt omgegaan als volgt verduidelijkt:

> "*When an approved blog links to another blog, that blog also gets approved after a while. But once in a while even real blogs link to fake blogs, which means that there will be a leak in the chain of trust. The spammy blog that gets approved will in turn approve potentially thousands of other spam blogs. When we detect this (i.e. through users reporting spam), we can follow the chain of trust upstream to find the faulting link and remove it. Which then disapproves the entire cluster of spam blogs.*" [99]

Aldus worden bij het ontdekken van spam in blog posts clusters van blog posts als geheel verwijderd.

Twingly hanteert standaard de TwinglyRank voor het ordenen van de blog posts. Wat de TwinglyRank is maakt de website van Twingly niet duidelijk. In een blog post op TechCrunch werd de TwinglyRank in april 2008 omgeschreven als:

> "*a combination of keyword relevancy, number of inlinks, number of user recommendations, publishing date and time and some secret sauce*" [100]

De TwinglyRank lijkt dus een combinatie van allerlei factoren.

In december 2008 introduceerde Twingly de BlogRank.[101] Deze ranking heeft niet betrekking op blog posts, maar op blogs. Volgens de FAQ de website van Twingly geeft deze BlogRank aan hoe populair een blog is binnen de taalgroep waartoe de blog behoort. Aanvullend merken de ontwikkelaars van Twingly in een blog post op dat de BlogRank een getal tussen 1 en 10 is, en dat deze BlogRank aangeeft hoe groot een blog is.[102] De ontwikkelaars vergelijken de BlogRank met de PageRank van Google, maar het verschil hiermee is dat de BlogRank alleen voor blogs is. Mijns inziens is het niet onwaarschijnlijk dat nu ook deze BlogRank mede de TwinglyRank bepaalt.

---

[99] <http://www.twingly.com/faq>
[100] Arrington 2008.
[101] Twingly 2008.
[102] Twingly 2008.

Technorati

Technorati biedt de mogelijkheid de gewenste autoriteitsgraad aan te geven. De autoriteitsgraad is een maatstaf waarin het aantal links van verschillende blogs naar een website gedurende de voorgaande 180 dagen tot uitdrukking komt.[103] Hierbij kan worden gekozen uit vier autoriteitsgraden[104], waarbij '*some authority*' de standaard graad is.

5.4.2 Verschillen ontsluiting blog posts

In paragraaf 3.5 kwam reeds aan de orde dat vooral blogzoekmachines nieuwe blog posts vinden aan de hand van de RSS-feeds van blogs. De mate waarin blogzoekmachines vervolgens de inhoud van blog posts ontsluiten, verschilt per zoekmachine.[105] Indien RSS-feeds niet de volledige inhoud van een blog post bevatten, halen Google Blog Search en Icerocket Blog Search de volledige blog post op en maken deze geheel doorzoekbaar. Technorati en Twingly Blog Search beperken zich tot de content van de RSS-feed. Deze laatste methode van ontsluiting kent als nadeel dat zij kan leiden tot een onevenwichtige weergave van blogs met en zonder volledige RSS-feed. Deze onevenwichtigheid kan ertoe leiden dat een relevante blog post geheel niet als zoekresultaat voorkomt.[106] Ook kan deze onevenwichtigheid ertoe leiden dat een blog post op een lagere positie als zoekresultaat wordt vermeld dan wanneer de volledige inhoud van de blog post in de RSS-feed was overgenomen.

5.5 Conclusie

Tussen de vijf zoekmachines bestaan grote verschillen in achtergrond en functionaliteiten. De belangrijkste verschillen tussen de blogzoekmachines onderling liggen op het vlak van sortering en de mate waarin de zoekmachines blog posts ontsluiten. Google Blog Search en Twingly Blog Search sorteren standaard op relevantie. Technorati en Icerocket Blog Search bieden resultaten slechts in omgekeerd chronologische volgorde aan. Ook op het vlak van de mate waarin blog posts worden ontsloten, kunnen twee groepen worden onderscheiden. Met Google Blog Search en Icerocket Blog Search zijn blog posts geheel doorzoekbaar; ook indien de RSS-feeds niet de volledige inhoud weergeven. Technorati en Twingly Blog Search beperken zich tot de inhoud van RSS-feeds van blogs.
De algemene internetzoekmachine Google Search verschilt van de blogzoekmachines, in het ontbreken van een betrouwbare mogelijkheid om webpagina's in omgekeerd chronologische volgorde te ordenen. In tegenstelling tot de blogzoekmachines doorzoekt Google Search gehele webpagina's, met inbegrip van reacties op blog posts.

---

[103] <http://support.technorati.com/faq/topic/71>
[104] Tot half februari 2009 kon nog worden gekozen uit vijf autoriteitsgraden.
[105] Vgl. tabel 2 en voetnoten 93 t/m 96.
[106] Vgl. reactie Matt Cutts in Gray 2009.

Ondanks de verschillen is het goed mogelijk de kwaliteit van de vijf zoekmachines door middel van een experiment te evalueren. Van alle vijf de zoekmachines mag namelijk worden verwacht dat de meest relevante resultaten bovenaan worden vermeld.

6 OPZET EXPERIMENT

6.1 Inleiding

Nu in de hoofdstukken 2 tot en met 5 de blogs, de zoekers van blogs en de zoekmachines voor blogs zijn onderzocht, wordt vervolgens toegekomen aan de volgende deelvraag: 'Hoe effectief zijn blogzoekmachines bij het vinden van relevante blog posts en relevante overige webpagina's, en wat is de toegevoegde waarde ten opzichte van algemene internetzoekmachines?'. Zoals in de inleiding vermeld wordt deze deelvraag over twee hoofdstukken verdeeld. In dit hoofdstuk wordt beschreven op welke wijze het experiment – of eigenlijk de drie deelexperimenten – wordt verricht. In het volgende hoofdstuk worden de resultaten van de deelexperimenten gepresenteerd en geanalyseerd.
Dit hoofdstuk is als volgt opgebouwd: in paragraaf 6.2 wordt allereerst beschreven wat met het experiment wordt beoogd te onderzoeken. Vervolgens wordt in paragraaf 6.3 beschreven waarom welke zoekmachines in de experiment worden betrokken. Hierna wordt in paragraaf 6.4 beschreven op welke wijze de zoekvragen worden opgesteld, en wordt in paragraaf 6.5 beschreven op welke wijze de zoekvragen worden uitgevoerd. In deze paragraaf wordt het experiment in drie deelexperimenten onderscheiden. Tot slot wordt in paragraaf 6.6 beschreven op welke wijze de gevonden resultaten worden geanalyseerd.

6.2 Doel experiment

Door het uitvoeren van het experiment dat in dit hoofdstuk beschreven wordt, kan een algemeen beeld worden verkregen van de retrieval effectiviteit van blogzoekmachines bij het vinden van relevante blog posts en relevante overige webpagina's. Tevens kan een algemeen beeld van de toegevoegde waarde van blogzoekmachines ten opzichte van algemene internetzoekmachines worden verkregen.

Uit de ordening binnen blogs kan worden opgemaakt dat recentheid inherent is aan het medium van de blog. De meest recente bijdragen staan namelijk altijd bovenaan vermeld. Alle zoekvragen en achterliggende informatiebehoeften beogen daarom informatie die niet ouder is dan drie maanden.[107]
Daarnaast zal een internetter via een blogzoekmachine niet snel zoekvragen over het persoonlijke leven van onbekenden uitvoeren. Daarom beogen alle zoekvragen en achterliggende informatiebehoeften slechts de niet of nauwelijks persoonlijke onderwerpen.
In verband met de beperkte tijd wordt in het experiment geen aandacht besteed aan het zoeken naar 'backlinks' en het zoeken naar blogs.

[107] Vgl. A. Agarwal 2009.

6.3 Motivering zoekmachines

Om te komen tot een algemeen beeld van de prestaties van blogzoekmachines worden de volgende vier toonaangevende blogzoekmachines in het experiment betrokken:

- Google Blog Search (NL)
- Icerocket
- Technorati
- Twingly

Icerocket en Twingly zijn beduidend minder populair dan Google Blog Search en Technorati (zie hoofdstuk 5), maar omdat de dekking van deze zoekmachines voldoende breed is, en omdat beide zoekmachines vernieuwingsgericht zijn, worden beide in het experiment betrokken.

Voor één van de drie deelexperimenten worden zoekvragen tevens uitgevoerd in de algemene zoekmachine van Google: Google Search (NL). Door de zoekvragen niet alleen in blogzoekmachines uit te voeren, maar ook in een algemene zoekmachine, kan de toegevoegde waarde van blogzoekmachines worden onderzocht. Als algemene zoekmachine is gekozen voor Google Search, omdat deze zoekmachine het meest wordt gebruikt in Nederland en veel andere landen.

6.4 Wijze van opstellen zoekvragen

In alle drie de deelexperimenten worden dezelfde zoekvragen uitgevoerd. Mede op basis van het in paragraaf 4.2 vermelde onderzoek naar de onderwerpen van blogs die worden gelezen, worden de volgende vijf categorieën onderwerpen van zoekvragen opgesteld:

- Internetzoekmachines
- Technologie (overig)
- New York (vanuit toeristisch oogpunt)
- Wielrennen (sport)
- Noord-Brabant (regionale ontwikkelingen)

Binnen iedere onderwerpscategorie wordt een tweede onderscheid gemaakt naar type van de zoekvraag:

- Personen
- Organisaties
- Producten
- Locaties
- Hogere concepten / onderwerpen

In de deelexperimenten wordt ervan uitgegaan dat een zoekvraag steeds in slechts één van de vijf categorieën kan worden ingedeeld. Voor iedere categorie (naar onderwerp en naar type) worden tien zoekvragen en achterliggende informatiebehoeften opgesteld. In Appendix A worden de vijftig zoekvragen met de categorieën waarbinnen de zoekvragen vallen weergegeven.

Bij het opstellen van de zoekvragen zijn op diverse gebieden keuzes gemaakt. Allereerst moet de gebruikte taal worden gemotiveerd. Indien zoekvragen niet uitsluitend uit namen bestaan, moet een taal waarin de zoekvragen worden opgesteld worden gekozen. In de deelexperimenten is ervoor gekozen om de zoekvragen in de categorieën internetzoekmachines, technologie, New York en wielrennen op te stellen in de Engelse taal. In deze categorieën verdient het Engels de voorkeur boven het Nederlands, omdat in het Engels meer over deze onderwerpen wordt geblogd. De zoekvragen in de categorie Noord-Brabant worden in de Nederlandse taal opgesteld. In deze categorie is het Nederlands als taal gekozen omdat het Nederlands in de provincie Noord-Brabant de gangbare taal is, en in het Nederlands beduidend meer blog posts over Noord-Brabant bestaan.

Ten tweede kan de lengte van de zoekvragen worden gemotiveerd. Overeenkomstig het in hoofdstuk 4 beschreven onderzoek zullen zoekvragen bij benadering tussen de twee en drie woorden bedragen. Aanhalingstekens worden niet gebruikt, omdat zoekmachines deze verschillend kunnen interpreteren en deze weinig worden gebruikt. Om dezelfde redenen worden Booleaanse operatoren (AND, OR, NOT en NEAR) niet gebruikt.

Voor het overgrote deel zullen van de namendragende entiteiten en onderwerpen zoveel mogelijk de zeer bekende worden gekozen. Bij het opstellen van deze namendragende entiteiten en onderwerpen wordt mede Wikipedia gehanteerd, omdat deze encyclopedie op uiteenlopende terreinen de belangrijkste namendragende entiteiten en onderwerpen bevat, en hun onderlinge samenhang weergeeft.[108]

[108] <http://en.wikipedia.org/>

Tot slot moet worden verduidelijkt dat in de deelexperimenten slechts recente informatie wordt beoogd.

Ter afsluiting van deze paragraaf volgt een voorbeeld van een zoekvraag en en achterliggende informatiebehoefte voor de categorie New York:

Zoekvraag: Affinia Manhattan
Achterliggende informatiebehoefte: Recente informatie of ervaringen ten aanzien van het Affinia Manhattan hotel.

6.5 Wijze van uitvoeren zoekvragen

Zoals reeds opgemerkt is eigenlijk niet sprake van één, maar van drie deelexperimenten:

- Met deelexperiment 1 wordt de effectiviteit van de vier blogzoekmachines onderzocht bij het vinden van relevante blog posts,
- Met deelexperiment 2 wordt de effectiviteit van de vier blogzoekmachines onderzocht bij het vinden van relevante blog posts en relevante overige webpagina's,
- Met deelexperimenten 3a en 3b wordt de effectiviteit van Google Blog Search (NL) en Google Search (NL) onderzocht: zowel bij het vinden van alleen relevante blog posts, als bij het vinden van relevante blog posts en relevante overige webpagina's.

Deelexperimenten 1 en 2 zijn tegelijkertijd uitgevoerd. De resultaten met betrekking tot Google Blog Search uit deelexperimenten 1 en 2 zijn gebruikt in deelexperimenten 3a en 3b. De vijftig zoekvragen zijn allen in de periode 8 tot en met 13 maart 2009 in de blogzoekmachines uitgevoerd. Daarbij zijn de afzonderlijke zoekvragen in korte tijd in de verschillende blogzoekmachines uitgevoerd. Aldus kunnen de resultaten van de blogzoekmachines zo goed mogelijk vergeleken worden. De vijftig zoekvragen zijn ongeveer 2 weken later in Google Search uitgevoerd.[109] Gelet op de verschillen in resultaten die zich in Google Search en Google Blog Search afhankelijk van de locatie van benadering kunnen voordoen, wordt opgemerkt dat alle zoekacties vanuit Zuid-Nederland zijn uitgevoerd.

De effectiviteit van de zoekmachines wordt in alle drie de deelexperimenten vastgesteld op basis van de top 10 van zoekresultaten die de zoekmachine biedt.[110] Mede gebaseerd op het onderzoek uit

[109] Mijns inziens vormt het verschil in periode geen probleem voor de vergelijkbaarheid tussen Google Blog Search en Google Search. Van Google Search zou mogen worden verwacht dat deze – over de vijftig zoekvragen genomen – over een periode van twee weken niet tot zeer verschillende resultaten leidt.
[110] Zie ook paragraaf 4.3.

paragraaf 4.3 wordt in de experimenten ervan uitgegaan dat iedere informatiebehoefte steeds slechts tot één zoekvraag en zoekactie leidt.

Voor ieder resultaat uit de top 10 wordt geoordeeld of het als relevant wordt beschouwd en of het resultaat een blog / blog post is of niet. Deze relevantie wordt beoordeeld vanuit de vooraf vastgestelde, achterliggende informatiebehoefte. Daarbij wordt aan ieder resultaat één van de in Tabel 3 weergegeven waarden toegekend:

Tabel 3 - Toegekende waarden

Waarde	Betekenis
0	Niet relevant
1	Geen blog / blog post, maar de webpagina is wel relevant
2	Relevante blog / blog post

Om zo objectief mogelijk te kunnen beoordelen of een zoekmachineresultaat relevant is of niet, en of de betreffende webpagina een blog of blog post is of niet, zijn beoordelingsregels opgesteld. Opgemerkt moet worden dat deze regels niet in alle gevallen tot een juiste beoordeling zullen leiden, omdat geen eenduidige criteria voor relevantie en blog bestaan.

Relevantie

Een zoekmachineresultaat wordt als relevant beschouwd indien in het resultaat de woorden van de zoekvraag (of een verbuiging of een vervoeging hiervan) in de context van de omschreven informatiebehoefte worden genoemd, en uit het resultaat een verband kan worden afgeleid tussen de woorden van deze zoekvraag en iets anders. Een zoekmachineresultaat wordt in de volgende gevallen echter als niet relevant beschouwd:

- In het resultaat komt alleen de vermelding zoals in de zoekvraag voor,
- Een resultaat bevat dezelfde inhoud als een bovenvermeld resultaat of een deel hiervan; naar elementen zoals lay-out en navigatie wordt hierbij niet gekeken,
- Een resultaat vermeldt in de URL of in de pagina een datum die ouder is dan drie maanden, of van een resultaat in een blogzoekmachine kan niet worden afgeleid wat de datum is,
- Een resultaat kan niet worden geladen,
- Het geheel van lay-out en tekst geeft evident de indruk dat het resultaat spam is.[111]

Blog / Blog post

Een zoekmachineresultaat wordt als blog of blog post beschouwd, indien het zoekmachineresultaat aan de definitie van blog uit hoofdstuk 2 voldoet. Aldus wordt als blog beschouwd:

[111] Deze conclusie kan bijvoorbeeld worden getrokken wanneer in extreme mate advertenties aanwezig zijn of wanneer de teksten onzinnig zijn. Indien een webpagina een (deel van een) post van een andere blog toont, en een verband tussen de woorden van de zoekvraag en iets anders kon worden afgeleid én deze originele blog post zelf niet erboven als zoekmachineresultaat voorkomt, dan wordt het zoekmachineresultaat toch als relevant beschouwd.

> “een type website dat in omgekeerd chronologische volgorde de bijdragen van één of meer individuen vermeldt”

Voor de praktische toetsbaarheid worden in het experiment aanvullend op deze definitie webpagina’s die voldoen aan de volgende kenmerken, niet als blog of blog post aangemerkt:

- de webpagina toont geen reacties en/of trackback links én biedt geen mogelijkheid tot het plaatsen van reacties en/of trackback links,
- de webpagina bevat een artikel van een persbureau, krant, tijdschrift, tv-station of radiozender zonder herkenbare toevoeging of aanpassing hiervan,
- de webpagina heeft de lay-out van een forum topic of forumoverzichtspagina én vermeldt tevens de term forum (of een synoniem hiervan in welke taal dan ook).

Overzichtspagina’s binnen blogs (denk aan een pagina met de eerste 150 woorden van alle blog posts in een bepaalde maand), worden als blog beschouwd.

Met betrekking tot deelexperimenten 3a en 3b moet worden opgemerkt dat voornamelijk (maar niet uitsluitend) in Google Search tevens resultaten kunnen worden getoond van webpagina’s die geen blog post zijn. In dergelijke gevallen kan de plaatsingsdatum vaak niet (voldoende betrouwbaar) worden bepaald. In dit onderzoek zijn alleen de zoekmachineresultaten die evident oud zijn als niet-relevant beschouwd. Aldus kan een webpagina die geen aanwijzingen bevat ten aanzien van de datum waarop de pagina is geplaatst of voor het laatst is bijgewerkt als relevant worden beschouwd. Een pagina waarop bijvoorbeeld ‘©2006’ staat wordt echter als evident oud, en hiermee als niet-relevant beschouwd.

Ten tweede moet met betrekking tot Google Search worden opgemerkt dat deze zoekmachine tussen de gerangschikte resultaten ook video’s en afbeeldingen plaatst. Ook deze gerangschikte video’s en afbeeldingen zijn op relevantie beoordeeld. Daarbij worden video’s en afbeeldingen als relevant beschouwd indien de informatiebehoefte hierop duidelijk herkenbaar in beeld is. Indien de pagina’s waarnaar de video’s of afbeeldingen linken evident ouder dan drie maanden zijn, dan worden deze video’s of afbeeldingen als niet relevant beschouwd.
In dit onderzoek worden Google Maps kaarten met daarnaast lokale bedrijfsresultaten niet tot de gerangschikte resultaten in Google Search gerekend. Een dergelijke kaart met lokale bedrijfsresultaten wordt mijns inziens namelijk altijd boven de gerangschikte resultaten getoond. Een kaart zonder lokale bedrijfsresultaten wordt wel op relevantie beoordeeld, omdat deze niet altijd boven de gerangschikte resultaten voorkomt, maar soms ook binnen de gerangschikte resultaten.
Indien blog- of nieuwsartikelen gegroepeerd tussen de gerangschikte resultaten in Google Search voorkomen, dan worden deze resultaten afzonderlijk op relevantie beoordeeld.

Sinds 2009 plaatst Google Search geregeld boven de top van zoekresultaten voor de uitgevoerde zoekvraag een top 2 van resultaten voor een andere zoekvraag. Ook deze zogenaamde “bedoelde u” resultaten worden op relevantie beoordeeld.
Indien een resultatenpagina in Google Search meer dan 10 resultaten vermeldt, dan worden alleen de bovenste 10 resultaten op relevantie beoordeeld.

Naast voorgaande beoordelingsregels, moeten ook enkele opmerkingen worden gemaakt over de wijze waarop de zoekmachines bij de zoekacties worden ingesteld. Allereerst moet worden opgemerkt dat zoekvragen zoveel mogelijk op standaard wijze in de zoekmachines uitgevoerd. Dit houdt bijvoorbeeld in dat resultaten worden gesorteerd volgens de standaard sorteerwijze van de zoekmachine (vergelijk hoofdstuk 5). Dit heeft tot gevolg dat bijvoorbeeld de zoekresultaten van Technorati in omgekeerd chronologische volgorde worden geordend en de resultaten van Google naar relevantie worden geordend. Dit verschil in ordening levert mijns inziens geen problemen op bij het vergelijken, omdat de makers van de zoekmachines slechts verschillende opvattingen hebben over de meest optimale presentatie van de zoekresultaten.

Een tweede opmerking betreft de autoriteitsgraad in Technorati. In deze zoekmachine kan gekozen worden uit vier autoriteitsgraden[112], waarbij ‘some authority’ de standaard graad is. In de twee deelexperimenten wordt uitgegaan van deze standaard graad. Hierdoor worden blog posts van blogs met een lagere autoriteitsgraad niet in de zoekresultaten weergegeven.

Een derde opmerking betreft het spam-free zoeken met Twingly. Met Twingly kan op twee manieren worden gezocht: standaard en spam-free. Bij spam-free zoeken staat aangegeven dat deze optie zich nog in bètafase bevindt. Twingly maakt geen duidelijke keuze in de wijze waarop wordt gezocht (standaard of spam-free).[113] Aldus kan moeilijk worden aangegeven welke de standaard optie van deze zoekmachine is. In deelexperimenten 1 en 2 worden de zoekvragen spam-free uitgevoerd.

Tot slot moet worden benadrukt dat het onderhavige onderzoek een momentopname is. Ieder moment kunnen nieuwe blogs in de zoekmachines worden toegevoegd. Ook kunnen zoekmachines inconsistenties bevatten of kan de programmatuur worden aangepast. In al deze gevallen kan een nieuw experiment tot zeer verschillende resultaten leiden. Thelwall en Hasler ondervonden bijvoorbeeld in 2006 tijdens een verkennend onderzoek naar blogzoekmachines dat Google Blog Search afhankelijk van de sorteerwijze resultaten al dan niet toonden, en dat de vermelding van het aantal gevonden resultaten door de pagina te herladen kon wijzigen.[114] Ook tijdens een onderzoek naar

[112] Tot half februari 2009 kon nog worden gekozen uit vijf autoriteitsgraden.
[113] Vanuit het startscherm kan meteen op beide manieren worden gezocht door op de betreffende knop te klikken. Wanneer de zoekvraag niet wordt verstuurd door te klikken, maar door op de enter-toets in te drukken, dan wordt de zoekvraag standaard uitgevoerd. Wanneer deze zoekvraag echter vanuit het geavanceerd zoeken scherm wordt uitgevoerd door op de enter toets te drukken, dan wordt de zoekvraag spam-free uitgevoerd.
[114] Vgl. Thelwall & Hasler 2007, p. 475.

de prestaties van zoekmachines bij in het Russisch opgestelde zoekvragen in 2007[115], werden inconsistenties in Google Search vastgesteld.

De data van alle drie de deelexperimenten zijn opgenomen in Appendix B. Per zoekvraag wordt aanvullend op de relevantiebeoordeling per zoekvraag in woorden een kwaliteitsindicatie en overige bijzonderheden genoteerd. Deze kwaliteitsindicatie en overige bijzonderheden worden van lagere waarde geschat, en daarom niet uitvoerig geanalyseerd.

6.6 Wijze van analyse resultaten

In paragraaf 3.5 werd een artikel van Tague-Sutcliffe aangehaald, dat als inspiratie heeft gediend voor dit experiment. De retrieval effectiviteit van de te onderzoeken zoekmachines wordt uitgedrukt in de maatstaven precisie en succes. Onder precisie wordt verstaan: het aandeel van het aantal relevante resultaten in het totale aantal resultaten voor een zoekvraag.[116] Onder het begrip precision@x wordt daarbij verstaan: het aandeel van het aantal relevante resultaten in de eerste x resultaten voor een zoekvraag. Uit de geregistreerde waarden, kunnen voor alle vijftig zoekvragen gezamenlijk de precision@1 t/m precision@10 worden afgeleid. Daarnaast kunnen precision@1 t/m precision@10 ook op tien wijzen voor steeds een vijfde van de vijftig zoekvragen worden berekend.
Naast de maatstaf precisie wordt de maatstaf succes gehanteerd. De maatstaf succes geeft aan of minimaal 1 relevant resultaat kan worden gevonden. Evenals bij de precisie kan succes worden bepaald naar positie in de zoekmachineresultaten. Over één zoekvraag berekend kan de maatstaf success@x twee waarden aannemen: 0 en 1. De waarde is 0 indien op positie 1 t/m x geen relevante resultaten kunnen worden gevonden. De waarde is 1 indien minimaal één resultaat relevant is.
Aanvullend op de maatstaven precisie en succes worden de totalen van geregistreerde waarden (0, 1 en 2) bestudeerd.

[115] Meijer 2007.
[116] Tague-Sutcliffe 1992, p. 472.

7 RESULTATEN EXPERIMENTEN

7.1 Inleiding

Nadat de drie deelexperimenten zijn uitgevoerd en geanalyseerd, kunnen de resultaten worden gepresenteerd. Allereerst worden in paragraaf 7.2 de resultaten van het onderzoek naar de effectiviteit van de vier blogzoekmachines bij het vinden van relevante blog posts (deelexperiment 1) gepresenteerd. Vervolgens worden in paragraaf 7.3 de resultaten van het onderzoek naar de effectiviteit van de vier blogzoekmachines bij het vinden van relevante blog posts en relevante overige webpagina's (deelexperiment 2) gepresenteerd. Hierna worden in paragraaf 7.4 de resultaten van het onderzoek naar Google Blog Search en Google Search (deelexperimenten 3a en 3b) gepresenteerd. In paragraaf 7.5 vindt vervolgens een analyse van de kwaliteit van de rangschikking plaats, waarna in paragraaf 7.6 de statistieken van alle deelexperimenten samen worden weergegeven. Paragraaf 7.7 sluit het hoofdstuk af met een conclusie.

7.2 Resultaten deelexperiment 1

Sommige mensen gebruiken een blogzoekmachine uitsluitend om blog posts te vinden. Deze mensen zijn in de blogzoekmachine bijvoorbeeld niet op zoek naar een webpagina met alleen een krantenartikel. Daarom wordt in deelexperiment 1 onderzocht hoe goed de retrieval effectiviteit van de blogzoekmachines is, indien alleen blog posts op relevantie worden beoordeeld, en overige webpagina's altijd als niet relevant worden beschouwd.

7.2.1 Resultaten m.b.t. totale aantal zoekvragen

Uit Tabel 4 kan worden afgeleid dat de retrieval effectiviteit van de vier blogzoekmachines bij het vinden van relevante blog posts over het totaal van vijftig zoekvragen in deelexperiment 1 uiteenloopt. Op basis van de maatstaven precision@1, precision@5 en precision@10 levert Twingly (spam-free) met afstand de beste prestaties. Gemiddeld zijn iets meer dan de helft van de resultaten relevant, terwijl een derde tot tweevijfde van de resultaten in de andere blogzoekmachines relevant is. Op basis van de maatstaf success@10 levert Google Blog Search (NL) de beste prestaties. Twingly geeft iets minder vaak minimaal één relevant resultaat in de top 10. Op basis van de maatstaf success@5 zijn de prestaties van Google Blog Search en Twingly gelijk. De zoekmachines Icerocket en Technorati presteren beduidend minder goed dan de twee andere blogzoekmachines. Onderling vergeleken liggen de prestaties van Icerocket licht hoger dan die van Technorati.

7.2.2 Resultaten m.b.t. categorieën

Uit Tabel 5 kan worden afgeleid dat de retrieval effectiviteit van de vier blogzoekmachines bij het vinden van blog posts in de tien categorieën in deelexperiment 1 uiteenloopt. De prestaties van Twingly blijven in de categorie Hogere concepten / onderwerpen en – in iets mindere mate – in de categorie Noord-Brabant beduidend achter bij de prestaties in de andere categorieën. Dit verschil is ook zichtbaar in de resultaten van Icerocket en Technorati, maar niet in Google Blog Search. In Google Blog Search liggen de resultaten in de categorie Hogere concepten / onderwerpen dicht bij de gemiddelden over het totaal van vijftig zoekvragen, en wijken alleen de prestaties in de categorie Noord-Brabant beduidend af.

Op basis van een aantal maatstaven kan worden geconcludeerd dat in de categorie Hogere concepten / onderwerpen de prestaties van Google Blog Search de prestaties van Twingly (en de twee andere blogzoekmachines) overstijgen. In de andere categorieën overstijgen de prestaties van Twingly de prestaties van Google Blog Search op basis van een groot aantal maatstaven.

Tabel 4 Resultaten deelexperiment 1 (alleen blog posts zijn relevant, totaal van 50 zoekvragen)

Google Blog Search (NL)					Technorati					Icerocket					Twingly (spam-free)				
p@1	**p@5**	**p@10**	**s@5**	**s@10**	**p@1**	**p@5**	**p@10**	**s@5**	**s@10**	**p@1**	**p@5**	**p@10**	**s@5**	**s@10**	**p@1**	**p@5**	**p@10**	**s@5**	**s@10**
0,42	0,42	0,42	0,82	0,90	0,30	0,34	0,32	0,58	0,62	0,38	0,36	0,34	0,68	0,80	0,62	0,56	0,54	0,82	0,84

N.B.: p@1 = s@1

Tabel 5 Resultaten deelexperiment 1 (alleen blog posts zijn relevant, verdeling over de 10 categorieën)

Google Blog Search (NL)					Technorati					Icerocket					Twingly (spam-free)					
p@1	**p@5**	**p@10**	**s@5**	**s@10**	**p@1**	**p@5**	**p@10**	**s@5**	**s@10**	**p@1**	**p@5**	**p@10**	**s@5**	**s@10**	**p@1**	**p@5**	**p@10**	**s@5**	**s@10**	
0,40	0,44	0,48	0,90	1,00	0,20	0,28	0,27	0,60	0,70	0,30	0,26	0,23	0,60	0,80	0,60	0,66	0,73	1,00	1,00	**1 Internetzoekmachines**
0,70	0,66	0,65	1,00	1,00	0,40	0,50	0,45	0,80	0,80	0,70	0,58	0,50	0,90	0,90	0,80	0,78	0,72	0,80	0,80	**2 Technologie (overig)**
0,30	0,58	0,53	1,00	1,00	0,30	0,40	0,41	0,60	0,60	0,40	0,32	0,34	0,60	0,80	0,60	0,64	0,63	0,90	1,00	**3 New York**
0,40	0,30	0,27	0,70	0,70	0,40	0,44	0,39	0,70	0,80	0,20	0,44	0,42	0,80	0,80	0,50	0,50	0,48	0,80	0,80	**4 Wielrennen**
0,30	0,14	0,19	0,50	0,80	0,20	0,10	0,07	0,20	0,20	0,30	0,18	0,19	0,50	0,70	0,60	0,22	0,13	0,60	0,60	**5 Noord-Brabant**
0,60	0,46	0,43	0,90	1,00	0,30	0,40	0,39	0,80	0,80	0,50	0,48	0,46	0,80	0,90	0,70	0,74	0,72	0,90	0,90	**A Personen**
0,20	0,36	0,34	0,70	0,70	0,30	0,26	0,21	0,50	0,60	0,30	0,32	0,30	0,70	0,80	0,70	0,52	0,44	0,90	0,90	**B Organisaties**
0,50	0,48	0,48	0,80	0,90	0,20	0,40	0,40	0,60	0,60	0,50	0,46	0,43	0,80	0,90	0,60	0,68	0,67	0,80	0,80	**C Producten**
0,40	0,44	0,47	0,90	1,00	0,60	0,60	0,52	0,90	0,90	0,30	0,42	0,38	0,80	0,80	0,90	0,68	0,65	1,00	1,00	**D Locaties**
0,40	0,38	0,40	0,80	0,90	0,10	0,06	0,07	0,10	0,20	0,30	0,10	0,11	0,30	0,60	0,20	0,18	0,21	0,50	0,60	**E Hogere concepten / onderwerpen**

N.B.: p@1 = s@1

7.3 Resultaten deelexperiment 2

Niet alle mensen zoeken met blogzoekmachines uitsluitend naar relevante blog posts. Sommige mensen vinden naast relevante blog posts ook graag relevante overige webpagina's (zoals pagina's met krantenartikelen). Daarom is aanvullend op deelexperiment 1 onderzocht hoe goed de retrieval effectiviteit van de blogzoekmachines is indien alle resultaten op relevantie worden beoordeeld.

7.3.1 Resultaten m.b.t. totale aantal zoekvragen

Uit Tabel 6 kan worden afgeleid dat de retrieval effectiviteit van de vier blogzoekmachines over het totaal van vijftig zoekvragen in deelexperiment 2 uiteenloopt. Op basis van alle vijf de maatstaven levert Google Blog Search de beste prestaties. Op basis van de maatstaven precision@1, precision@5 en precision@10 beoordeeld presteert Twingly beduidend minder goed. Icerocket volgt kort achter Twingly, en tot slot volgt Technorati op een grotere afstand achter Icerocket. Op basis van de success@5 en success@10 maatstaven volgen Twingly en Icerocket kort achter Google Blog Search. Technorati volgt op beduidend grote afstand.

7.3.2 Resultaten m.b.t. categorieën

Uit Tabel 7 kan worden afgeleid dat de retrieval effectiviteit van de vier blogzoekmachines in de tien categorieën in deelexperiment 2 uiteenloopt. Google Blog Search en Twingly scoren in de categorieën Internetzoekmachines, Technologie (overig), New York en Locaties redelijk gelijk. Google Blog Search scoort in de categorieën Wielrennen, Noord-Brabant, Organisaties, Producten en Hogere concepten / onderwerpen iets hoger of beduidend hoger dan Twingly. In de categorie Noord-Brabant scoort Icerocket beduidend hoger dan Google Blog Search. Tot slot scoort Twingly in de categorie Personen licht hoger dan Google Blog Search.

Tabel 6 Resultaten deelexperiment 2 (alle webpagina's die in de informatiebehoefte voorzien worden als relevant beschouwd, totaal van 50 zoekvragen)

Google Blog Search (NL)					Technorati					Icerocket					Twingly (spam-free)				
p@1	**p@5**	**p@10**	**s@5**	**s@10**	**p@1**	**p@5**	**p@10**	**s@5**	**s@10**	**p@1**	**p@5**	**p@10**	**s@5**	**s@10**	**p@1**	**p@5**	**p@10**	**s@5**	**s@10**
0,74	0,68	0,67	0,98	1,00	0,42	0,42	0,38	0,68	0,72	0,60	0,51	0,48	0,76	0,84	0,68	0,59	0,57	0,82	0,84

N.B.: p@1 = s@1

Tabel 7 Resultaten deelexperiment 2 (alle webpagina's die in de informatiebehoefte voorzien worden als relevant beschouwd, verdeling over de 10 categorieën)

Google Blog Search (NL)					Technorati					Icerocket					Twingly (spam-free)					
p@1	**p@5**	**p@10**	**s@5**	**s@10**	**p@1**	**p@5**	**p@10**	**s@5**	**s@10**	**p@1**	**p@5**	**p@10**	**s@5**	**s@10**	**p@1**	**p@5**	**p@10**	**s@5**	**s@10**	
0,70	0,60	0,60	1,00	1,00	0,20	0,30	0,28	0,70	0,80	0,40	0,30	0,25	0,60	0,80	0,60	0,66	0,73	1,00	1,00	**1 Internetzoekmachines**
0,90	0,82	0,80	1,00	1,00	0,50	0,58	0,51	0,90	0,90	0,80	0,66	0,58	0,90	0,90	0,80	0,80	0,77	0,80	0,80	**2 Technologie (overig)**
0,60	0,74	0,67	1,00	1,00	0,40	0,42	0,43	0,60	0,60	0,50	0,36	0,37	0,60	0,80	0,70	0,68	0,68	0,90	1,00	**3 New York**
0,70	0,70	0,74	1,00	1,00	0,50	0,56	0,52	0,70	0,80	0,40	0,52	0,56	0,80	0,80	0,70	0,60	0,53	0,80	0,80	**4 Wielrennen**
0,80	0,54	0,54	0,90	1,00	0,50	0,24	0,16	0,50	0,50	0,90	0,72	0,63	0,90	0,90	0,60	0,22	0,14	0,60	0,60	**5 Noord-Brabant**
0,80	0,76	0,72	1,00	1,00	0,50	0,46	0,46	0,90	0,90	0,70	0,64	0,60	0,80	0,90	0,90	0,80	0,76	0,90	0,90	**A Personen**
0,60	0,64	0,62	1,00	1,00	0,30	0,40	0,29	0,60	0,70	0,50	0,52	0,48	0,80	0,80	0,70	0,56	0,48	0,90	0,90	**B Organisaties**
0,90	0,72	0,74	0,90	1,00	0,30	0,46	0,48	0,60	0,60	0,60	0,54	0,50	0,80	0,90	0,70	0,72	0,70	0,80	0,80	**C Producten**
0,80	0,72	0,69	1,00	1,00	0,60	0,62	0,54	0,90	0,90	0,70	0,62	0,59	0,90	0,90	0,90	0,68	0,68	1,00	1,00	**D Locaties**
0,60	0,56	0,58	1,00	1,00	0,40	0,16	0,13	0,40	0,50	0,50	0,24	0,22	0,50	0,70	0,20	0,20	0,23	0,50	0,60	**E Hogere concepten / onderwerpen**

N.B.: p@1 = s@1

7.4 Resultaten deelexperimenten 3a en 3b

Sommige mensen geven aan nooit zoekmachines te gebruiken. Deze mensen zoeken veelal met een algemene zoekmachine. In deelexperimenten 3a en 3b is vergeleken in hoeverre de retrieval effectiviteit van Google Search (NL) verschilt van Google Blog Search (NL).

7.4.1 Resultaten deelexperiment 3a

Uit Tabel 8 kan worden afgeleid dat de retrieval effectiviteit van Google Search en Google Blog Search over het totaal van vijftig zoekvragen in deelexperiment 3a fors uiteenloopt, indien alleen blog posts op relevantie worden beoordeeld (en andere webpagina's dus als niet relevant worden beschouwd). Google Search toont in tegenstelling tot Google Blog Search nauwelijks relevante resultaten. Gemiddeld is minder dan één resultaat binnen de top 10 relevant. Slechts voor een derde van de vijftig zoekvragen toonde Google Search minimaal één relevant resultaat in de top 10. Geen enkele keer toonde Google Search op positie 1 een relevante blog.

Tabel 8 Resultaten deelexperiment 3a (alleen blog posts zijn relevant, totaal van 50 zoekvragen)

Google Blog Search (NL)					Google Search (NL)				
p@1	**p@5**	**p@10**	**s@5**	**s@10**	**p@1**	**p@5**	**p@10**	**s@5**	**s@10**
0,42	0,42	0,42	0,82	0,90	0,00	0,05	0,06	0,16	0,34

N.B.: p@1 = s@1

Uit Tabel 9 kan worden afgeleid dat de retrieval effectiviteit van Google Search en Google Blog Search in de tien categorieën in experiment 3a fors uiteenloopt, indien alleen blog posts op relevantie worden beoordeeld. Slechts in de categorieën Internetzoekmachines, Producten en (in iets mindere mate) Technologie (overig) is de achterstand van Google Search op Google Blog Search iets minder fors, maar nog altijd duidelijk aanwezig.

Tabel 9 Resultaten experiment 3a (alleen blog posts zijn relevant, verdeling over de 10 categorieën)

Google Blog Search (NL)					Google Search (NL)					
p@1	**p@5**	**p@10**	**s@5**	**s@10**	**p@1**	**p@5**	**p@10**	**s@5**	**s@10**	
0,40	0,44	0,48	0,90	1,00	0,00	0,14	0,14	0,30	0,50	**1 Internetzoekmachines**
0,70	0,66	0,65	1,00	1,00	0,00	0,08	0,09	0,40	0,50	**2 Technologie (overig)**
0,30	0,58	0,53	1,00	1,00	0,00	0,00	0,05	0,00	0,30	**3 New York**
0,40	0,30	0,27	0,70	0,70	0,00	0,02	0,03	0,10	0,30	**4 Wielrennen**
0,30	0,14	0,19	0,50	0,80	0,00	0,00	0,01	0,00	0,10	**5 Noord-Brabant**
0,60	0,46	0,43	0,90	1,00	0,00	0,02	0,03	0,10	0,30	**A Personen**
0,20	0,36	0,34	0,70	0,70	0,00	0,04	0,06	0,20	0,30	**B Organisaties**
0,50	0,48	0,48	0,80	0,90	0,00	0,12	0,16	0,20	0,50	**C Producten**
0,40	0,44	0,47	0,90	1,00	0,00	0,04	0,03	0,20	0,30	**D Locaties**
0,40	0,38	0,40	0,80	0,90	0,00	0,02	0,04	0,10	0,30	**E Hogere concepten / onderwerpen**

N.B.: p@1 = s@1

7.4.2 Resultaten deelexperiment 3b

Uit Tabel 10 kan worden afgeleid dat de retrieval effectiviteit van Google Search en Google Blog Search over het totaal van vijftig zoekvragen in deelexperiment 3b redelijk gelijk is, indien resultaten op relevantie worden beoordeeld.[117] Indien slechts naar positie 1 wordt gekeken, dan biedt Google Search iets vaker relevante webpagina's. Op basis van de precision@10 maatstaf presteert Google Blog Search licht beter dan Google Search.

Tabel 10 Resultaten deelexperiment 3b (alle webpagina's die in de informatiebehoefte voorzien worden als relevant beschouwd, totaal van 50 zoekvragen)

Google Blog Search (NL)					Google Search (NL)				
p@1	**p@5**	**p@10**	**s@5**	**s@10**	**p@1**	**p@5**	**p@10**	**s@5**	**s@10**
0,74	0,68	0,67	0,98	1,00	0,82	0,66	0,56	0,96	0,98

N.B.: p@1 = s@1

Uit Tabel 11 kan worden afgeleid dat de retrieval effectiviteit van Google Search en Google Blog Search in de tien categorieën in deelexperiment 3b uiteenloopt, indien alle webpagina's die in de informatiebehoefte voorzien als relevant worden beschouwd. In de categorie Technologie (overig) presteert Google Blog Search aanmerkelijk beter dan Google Search. In de categorie Noord-Brabant presteert Google Search aanmerkelijk beter. In de categorieën Wielrennen en Personen presteert

[117] Nogmaals zij opgemerkt dat webpagina's in Google Search lastig kunnen worden onderscheiden in webpagina's jonger dan 3 maanden oud en vanaf 3 maanden oud. Vergelijk paragraaf 6.5.

Google Search voorin de top 10 beter. Middenin en achter in de top 10 scoort Google Blog Search beter. In de overige categorieën liggen de prestaties dichter bij elkaar.

Tabel 11 Resultaten deelexperiment 3b (alle webpagina's die in de informatiebehoefte voorzien worden als relevant beschouwd, verdeling over de 10 categorieën)

Google Blog Search (NL)					Google Search (NL)					
p@1	**p@5**	**p@10**	**s@5**	**s@10**	**p@1**	**p@5**	**p@10**	**s@5**	**s@10**	
0,70	0,60	0,60	1,00	1,00	0,80	0,62	0,47	0,90	0,90	**1 Internetzoekmachines**
0,90	0,82	0,80	1,00	1,00	0,70	0,52	0,44	0,90	1,00	**2 Technologie (overig)**
0,60	0,74	0,67	1,00	1,00	0,80	0,70	0,64	0,90	1,00	**3 New York**
0,70	0,70	0,74	1,00	1,00	0,80	0,62	0,53	1,00	1,00	**4 Wielrennen**
0,80	0,54	0,54	0,90	1,00	1,00	0,82	0,71	1,00	1,00	**5 Noord-Brabant**
0,80	0,76	0,72	1,00	1,00	0,90	0,66	0,51	1,00	1,00	**A Personen**
0,60	0,64	0,62	1,00	1,00	0,80	0,60	0,62	0,90	0,90	**B Organisaties**
0,90	0,72	0,74	0,90	1,00	0,90	0,68	0,63	1,00	1,00	**C Producten**
0,80	0,72	0,69	1,00	1,00	0,90	0,80	0,59	1,00	1,00	**D Locaties**
0,60	0,56	0,58	1,00	1,00	0,60	0,54	0,44	0,90	1,00	**E Hogere concepten / onderwerpen**

N.B.: p@1 = s@1

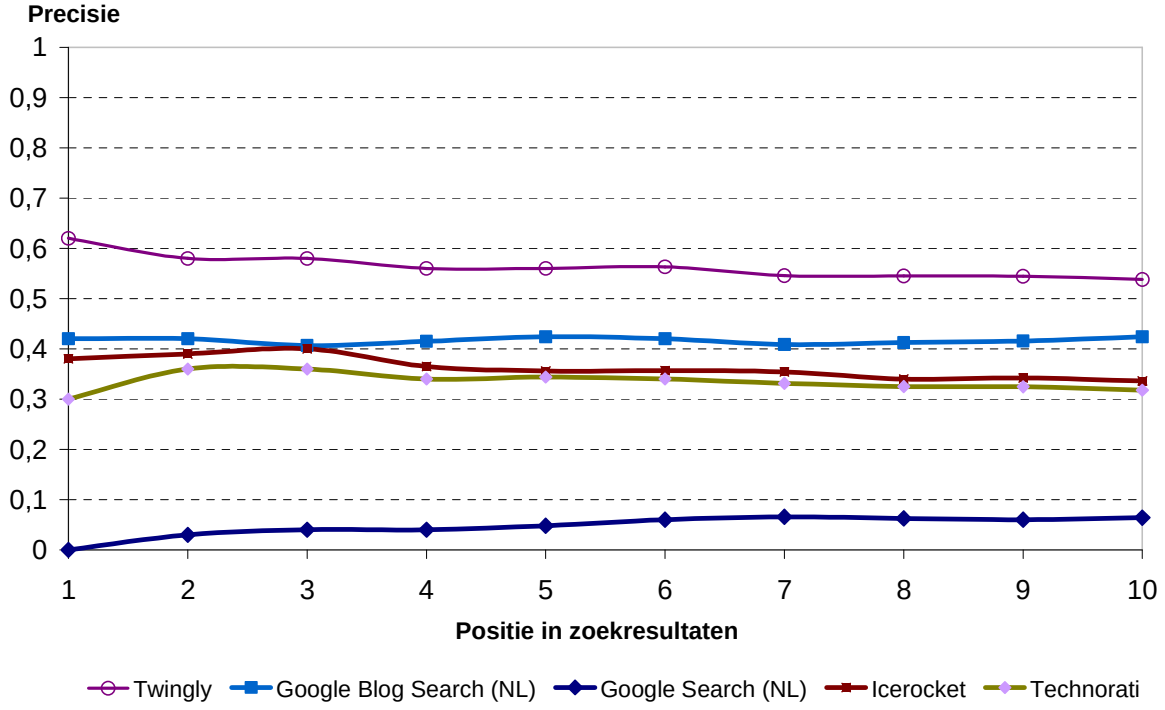
7.5 Kwaliteit rangschikking

Aan de hand van figuren van de precision@n maatstaf kunnen verschillen in precisie tussen de top 10 posities in de zoekresultaten zichtbaar worden gemaakt. Indien de precisie van een zoekmachine een licht dalende lijn vertoont, dan kan worden geconcludeerd dat de zoekmachine de meest relevante resultaten bovenin de top 10 plaatst, en duidelijk sprake is van een rangschikking. Indien de precisie een horizontale lijn of een stijgende lijn vertoont, dan kan worden geconcludeerd dat relevante resultaten niet altijd bovenaan in zoekresultaten van de zoekmachine voorkomen, maar net zo vaak of vaker middenin of achter in de top 10 van zoekresultaten.

In Figuur 1 wordt de precision@n maatstaf weergegeven voor de vijf onderzochte zoekmachines, indien alleen blog posts als relevant worden beschouwd. Op basis van de figuur kan worden geconcludeerd dat alleen de precisie van Twingly een licht dalende lijn laat zien. Aldus lijkt alleen bij Twingly sprake te zijn van een rangschikking: de resultaten zijn naar mate van relevantie gerangschikt. De precisie van Icerocket laat tot positie 3 een licht stijgende lijn zien, om vervolgens vanaf positie 3 te gaan dalen. De precisie van Technorati vertoont tot positie 2 een sterk stijgende lijn, om vervolgens geleidelijk te gaan dalen. De precisie van Google Blog Search vertoont een vrijwel constante lijn. Aldus kan worden geconcludeerd dat een zeer relevant resultaat zich binnen Google

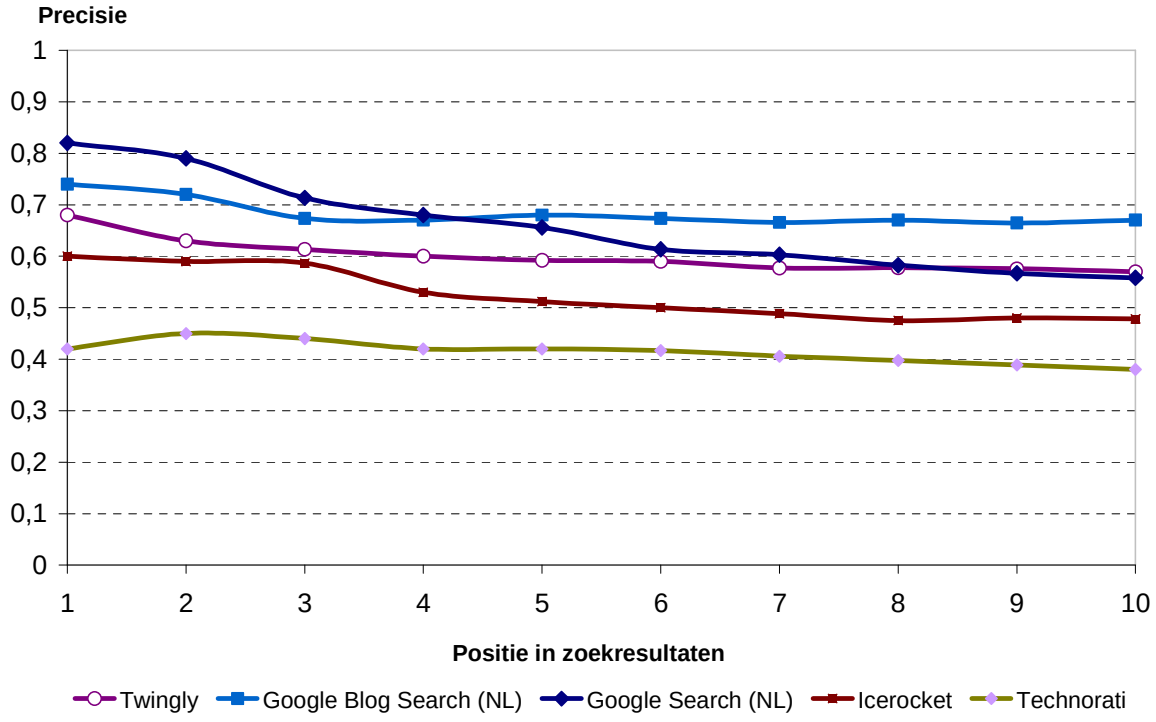
Blog Search op iedere positie binnen de top 10 kan bevinden. De precisie van Google Search vertoont een licht stijgende lijn. Aldus kan worden geconcludeerd dat relevante resultaten vaker middenin en achter in de top 10 worden weergegeven, dan bovenin de top 10 van zoekresultaten.

Figuur 1 – Precisie over het totaal van 50 zoekvragen (alleen blog posts worden als relevant beschouwd)



In Figuur 2 wordt de precision@n maatstaf weergegeven voor de vijf onderzochte zoekmachines, indien alle webpagina's die in de informatiebehoefte voorzien als relevant worden beschouwd. De precisie van Google Search vertoont de sterkst dalende lijn. De precisie van Twingly daalt bovenin de top 10 het sterkst, om vervolgens met een licht dalende lijn te vervolgen. De precisie van Google Blog Search daalt tot positie 3 fors, om hierna als een bijna horizontale lijn te vervolgen. De precisie van Icerocket blijft tot positie 3 constant, om hierna als een dalende lijn te vervolgen. De precisie van Technorati stijgt tot positie 2, om hierna als een licht dalende lijn te vervolgen. Gelet op de forse daling van de precisie van Google Search kan worden geconcludeerd dat in Google Search duidelijk sprake is van een rangschikking waarin de meest relevante resultaten op de hoogste posities worden weergegeven. In Google Blog Search en Twingly is deze rangschikking ook aanwezig, maar minder sterk. Tot slot is de rangschikking in Icerocket en Technorati slechts gedeeltelijk en tevens minder sterk aanwezig.

Figuur 2 – Precisie over het totaal van 50 zoekvragen (alle webpagina's die in de informatiebehoefte voorzien worden als relevant beschouwd)



Concluderend kan worden gesteld dat de kwaliteit van de rangschikking van Google Search verschilt afhankelijk van de vraag of alleen blog posts op relevantie worden beoordeeld, of dat alle resultaten op relevantie worden beoordeeld. In het eerste geval bevinden de meeste relevante blog posts zich achter in de top 10. Alleen in het tweede geval is het aan te bevelen Google Search te gebruiken. Daarbij moet worden bedacht dat de relevantie van ieder volgend resultaat in de top 10 fors minder is.

Daarnaast kan worden geconcludeerd dat Google Blog Search nauwelijks een rangschikking lijkt te kennen. Dit geldt zowel wanneer uitsluitend blog posts als relevant worden beschouwd, als wanneer alle webpagina's die in de informatiebehoefte voorzien als relevant worden beschouwd. Het is aan te bevelen in Google Blog Search niet alleen de eerste resultaten van de top 10 te bekijken, maar ook de lager gerangschikte resultaten te bekijken.

De licht dalende lijn die Twingly in beide figuren laat zien, duidt op de aanwezigheid van een goede rangschikking.[118] De rangschikking van Icerocket en Technorati kan tot slot als matig worden beschouwd.

[118] De kwaliteit van de rangschikking van Twingly lijkt mede te kunnen worden verklaard door het onderscheid dat Twingly maakt. Eerst worden alle relevante resultaten uit de voorgaande twee maanden gerangschikt. Pas nadat deze allemaal zijn gerangschikt, worden de resultaten ouder dan twee maanden gerangschikt. Vergelijk specifieke zoekvragen zoals <http://www.twingly.com/search?q=crisis+Nederland+%22wouter+bos%22>. In Google Blog Search en Google Search kon dit patroon niet worden gevonden. In Technorati en Icerocket was het patroon aanwezig door de automatische ordening naar datum.

7.6 Statistieken alle deelexperimenten

Tot slot kunnen in een tabel de totalen van alle geregistreerde waarden (0, 1 en 2) voor alle deelexperimenten samen worden weergegeven. Deze totalen staan weergegeven in Tabel 12.

Tabel 12 Statistieken alle deelexperimenten

	Google Search (NL)	Google Blog Search (NL)	Technorati	Icerocket	Twingly (spam-free)
Aantal 0	221	165	310	261	215
Aantal 1	247	123	31	71	16
Aantal 2	32	212	159	168	269

0 = niet relevant; 1 = geen blog / blog post, wel relevant; 2 = relevante blog / blog post

Uit Tabel 12 kan worden afgeleid dat Technorati het vaakst niet relevante resultaten weergaf. Google Blog Search (NL) gaf het minst vaak niet relevante resultaten weer. Technorati en Twingly (spam-free) toonden nauwelijks webpagina's die geen blog of blog post zijn. Tot slot toonde Google Search (NL) nauwelijks blogs of blog posts.

7.7 Conclusie

Aldus kan de vraag 'Hoe effectief zijn blogzoekmachines bij het vinden van relevante blog posts en relevante overige webpagina's, en wat is de toegevoegde waarde ten opzichte van algemene internetzoekmachines?' als volgt worden beantwoord:

Indien slechts blog posts op relevantie worden beoordeeld (en andere webpagina's dus als niet relevant worden beschouwd), dan is gemiddeld 1/3 tot 1/2 van de resultaten in de top 10 van een blogzoekmachine relevant. Afhankelijk van de zoekmachine beschikt een top 10 resultatenpagina gemiddeld in 60 tot 90 procent van de zoekvragen over ten minste één relevant resultaat. Onderling vergeleken levert Twingly (spam-free) de meeste relevante resultaten. Met uitzondering van Google Blog Search (NL) bieden de blogzoekmachines beduidend minder relevante resultaten voor de zoekvragen in de categorie Hogere concepten / onderwerpen.

Indien alle resultaten op relevantie worden beoordeeld, dan is gemiddeld 2/5 tot 7/10 van de resultaten in de top 10 van een blogzoekmachine relevant. Afhankelijk van de zoekmachine beschikt een top 10 resultatenpagina gemiddeld in 70 tot 100 procent van de zoekvragen over ten minste één relevant resultaat. Onderling vergeleken levert Google Blog Search (NL) de meeste relevante resultaten.

Indien de resultaten van Google Blog Search (NL) worden vergeleken met de algemene internetzoekmachine Google Search (NL), dan kan het volgende worden geconcludeerd. Indien alleen blog posts op relevantie worden beoordeeld (en andere webpagina's dus als niet relevant worden beschouwd), dan scoort Google Search (NL) zeer slecht. Gemiddeld is minder dan één resultaat binnen de top 10 relevant. Een top 10 resultatenpagina beschikt gemiddeld in 1/3 van de zoekvragen

over ten minste één relevant resultaat. Indien alle webpagina's op relevantie worden beoordeeld, dan scoren Google Search en Google Blog Search redelijk gelijk.

Over de kwaliteit van de rangschikking van de vijf zoekmachines kan worden opgemerkt dat alleen Twingly (spam-free) een echte rangschikking lijkt te kennen, indien alleen blog posts op relevantie worden beoordeeld. Relevante resultaten in Google Blog Search (NL) en Google Search (NL) bevinden zich vaak midden in of achter in de top 10. Indien alle webpagina’s die in de informatiebehoefte voorzien op relevantie worden beoordeeld, dan beschikt niet alleen Twingly (spam-free), maar ook Google Search (NL) over een echte rangschikking. Google Blog Search (NL) lijkt ook in dit geval nauwelijks over een rangschikking te beschikken.

8 CONCLUSIE

In hoofdstuk 1 werd de volgende probleemstelling opgesteld:

Hoe effectief zijn blogzoekmachines?

In hoofdstuk 2 werd allereerst het begrip blog gedefinieerd. In dit onderzoek wordt als blog beschouwd: het type website dat in omgekeerd chronologische volgorde de bijdragen van één of meer individuen vermeldt. Blog posts worden beschouwd als de bijdragen op blogs.
In hoofdstuk 3 werd vervolgens een theoretisch kader van de ontsluiting van informatie in het algemeen, en blogs in het bijzonder opgesteld. Geconcludeerd werd dat drie onderdelen binnen een ontsluitingssysteem een belangrijke rol spelen: de aanwezige informatie, de gebruikers en het systeem dat de informatie en de informatiebehoeften van de gebruikers matcht. Vooral op het gebied van de informatie verschilt de blogzoekmachine van de algemene internetzoekmachine. Blogzoekmachines richten zich op blogs, die beschikken over een structuur waarin de blog posts in omgekeerd chronologische volgorde geordend zijn, terwijl algemene internetzoekmachines zich op allerlei soorten webpagina’s richten. Ten aanzien van de verschillen tussen blogzoekmachines en algemene internetzoekmachines op het gebied van de gebruikers en het matchende systeem bestaat nauwelijks wetenschappelijk onderzoek.
In hoofdstuk 4 werd vervolgens onderzocht waarnaar en hoe internetters met blogzoekmachines zoeken. Geconcludeerd werd dat internetters blogzoekmachines vooral voor informationele zoekvragen te gebruiken. Daarbij zijn internetters die deze zoekmachines gebruiken vooral op zoek naar informatie over de zeer bekende personen en organisaties, en in iets minder mate hogere concepten en onderwerpen. Blogs over de onderwerpen muziek, nieuws / actuele zaken, meningen over producten/merken, film/tv, computers, reizen en technologie lijken internetters het vaakst te lezen. Over de wijze waarop internetters met blogzoekmachines zoeken kan worden afgeleid dat zij een aantal jaren terug gemiddeld twee tot drie woorden ingaven, meestal slechts de top 10 resultatenpagina opvroegen en meestal slechts één zoekvraag uitvoerden.
In hoofdstuk 5 werd vervolgens onderzocht wat de belangrijkste verschillen en overeenkomsten tussen de functionaliteiten van de blogzoekmachines onderling, en ten opzichte van Google Search zijn. Geconcludeerd werd dat grote verschillen tussen de zoekmachines bestaan. De belangrijkste verschillen tussen de blogzoekmachines onderling liggen op het vlak van sortering en de mate waarin de zoekmachines blog posts doorzoekbaar maken. De algemene internetzoekmachine Google Search verschilt van de blogzoekmachines, in het ontbreken van een betrouwbare mogelijkheid om webpagina's in omgekeerd chronologische volgorde te ordenen. In tegenstelling tot de blogzoekmachines doorzoekt Google Search gehele webpagina's, met inbegrip van reacties op blog posts.

In hoofdstuk 6 en 7 werden vervolgens de opzet en de resultaten van het onderzoek van de retrieval effectiviteit van de blogzoekmachines gepresenteerd. In het experiment werden vijftig zoekvragen in de blogzoekmachines Google Blog Search (NL), Technorati, Icerocket Blog Search en Twingly (spam-free) en in de algemene internetzoekmachine Google Search (NL) uitgevoerd. In deelexperiment 1 werden alleen de blog posts die niet ouder dan drie maanden zijn op relevantie beoordeeld. In deelexperiment 2 werden tevens andere webpagina's op relevantie beoordeeld. In deelexperiment 3 werd de retrieval effectiviteit van Google Blog Search (NL) en Google Search (NL) onderling onderzocht.
Op basis van de drie deelexperimenten kan worden geconcludeerd dat de toegevoegde waarde van blogzoekmachines afhankelijk is van de vraag of de internetter naast blog posts ook andere webpagina's die in de informatiebehoefte voorzien, getoond wil krijgen. Zo ja, dan is de toegevoegde waarde beperkt. Zo nee, dan hebben sommige blogzoekmachines duidelijk toegevoegde waarde ten opzichte van algemene internetzoekmachines.

Indien slechts blog posts op relevantie worden beoordeeld (en andere webpagina's dus als niet relevant worden beschouwd), dan is gemiddeld 1/3 tot 1/2 van de resultaten in de top 10 van een blogzoekmachine relevant. Afhankelijk van de zoekmachine beschikt een top 10 resultatenpagina gemiddeld in 60 tot 90 procent van de zoekvragen over ten minste één relevant resultaat. Onderling vergeleken levert Twingly (spam-free) de meeste relevante resultaten. Met uitzondering van Google Blog Search (NL) bieden de blogzoekmachines beduidend minder relevante resultaten voor de zoekvragen in de categorie Hogere concepten / onderwerpen.
Indien alle zoekmachineresultaten op relevantie worden beoordeeld, dan is gemiddeld 2/5 tot 7/10 van de resultaten in de top 10 van een blogzoekmachine relevant. Afhankelijk van de zoekmachine beschikt een top 10 resultatenpagina gemiddeld in 70 tot 100 procent van de zoekvragen over ten minste één relevant resultaat. Onderling vergeleken levert Google Blog Search (NL) de meeste relevante resultaten.
Indien de resultaten van Google Blog Search (NL) worden vergeleken met de algemene internetzoekmachine Google Search (NL), dan kan het volgende worden geconcludeerd. Indien alleen blog posts op relevantie worden beoordeeld (en andere webpagina's dus als niet relevant worden beschouwd), dan scoort Google Search (NL) zeer slecht. Gemiddeld is minder dan één resultaat binnen de top 10 relevant. Een top 10 resultatenpagina beschikt gemiddeld in 1/3 van de zoekvragen over ten minste één relevant resultaat. Indien alle webpagina's op relevantie worden beoordeeld, dan scoren Google Search en Google Blog Search redelijk gelijk.

Aldus kan de probleemstelling als volgt worden beantwoord: gekeken naar de retrieval effectiviteit bieden sommige blogzoekmachines duidelijk toegevoegde waarde ten opzichte van algemene internetzoekmachines, indien de internetter op zoek is naar relevante blog posts. Twingly (spam-free) levert de hoogste retrieval effectiviteit, terwijl Google Search (NL) de laagste retrieval effectiviteit

levert. Indien de internetter ook andere webpagina's die in de informatiebehoefte voorzien getoond wil krijgen, dan is toegevoegde waarde beperkt. In dit geval is de retrieval effectiviteit van Google Blog Search (NL) en Google Search (NL) het hoogste.

De vraag kan worden gesteld in hoeverre de resultaten van het experiment representatief zijn voor andere taalversies van Google Search en Google Blog Search. Naar mijn indruk presenteren de Nederlandse versies veel Nederlandstalige resultaten in de top 10. In andere taalversies zal Google deze resultaten waarschijnlijk minder hoog vermelden. De locatie van waaruit de zoekvragen worden uitgevoerd, kan mogelijk ook tot verschillen leiden. Mijns inziens kan een experiment met een andere taalversie of locatie daarom tot verschillende resultaten leiden. Tevens moet worden bedacht dat blogzoekmachines voortdurend bezig zijn hun systemen te optimaliseren. Ook dit zou bij een nieuw experiment kunnen leiden tot verschillende resultaten. Nader onderzoek op deze gebieden is zeer gewenst.

APPENDIX A: ZOEKVRAGEN MET INDELING IN CATEGORIEEN

Zoekvraag	Categorie (onderwerp)	Categorie (type)
Sergey Brin	Internetzoekmachines	Personen
Jerry Yang	Internetzoekmachines	Personen
Google corporate	Internetzoekmachines	Organisaties
Baidu english language search	Internetzoekmachines	Organisaties
Twingly Microblog	Internetzoekmachines	Producten
FriendFeed Search	Internetzoekmachines	Producten
Mountain View	Internetzoekmachines	Locaties
Sunnyvale	Internetzoekmachines	Locaties
Google new features	Internetzoekmachines	Hogere concepten / onderwerpen
search engines developments	Internetzoekmachines	Hogere concepten / onderwerpen
Steve Jobs	Technologie (overig)	Personen
Steve Ballmer	Technologie (overig)	Personen
Adobe	Technologie (overig)	Organisaties
Mozilla open-source	Technologie (overig)	Organisaties
Yahoo Pipes mashups	Technologie (overig)	Producten
Firefox bugs	Technologie (overig)	Producten
Silicon Valley	Technologie (overig)	Locaties
Bangalore technology	Technologie (overig)	Locaties
web 4.0	Technologie (overig)	Hogere concepten / onderwerpen
technology sector financial survival	Technologie (overig)	Hogere concepten / onderwerpen
Michael Bloomberg	New York	Personen
Frederick Law Olmsted	New York	Personen
NYC Information Center	New York	Organisaties
Affinia Manhattan	New York	Organisaties
meals New York	New York	Producten
MTA fares	New York	Producten
Central Park NYC	New York	Locaties
Grand Central Terminal	New York	Locaties
literature New York	New York	Hogere concepten / onderwerpen
sightseeing New York	New York	Hogere concepten / onderwerpen
Lance Armstrong	Wielrennen	Personen
Pat McQuaid	Wielrennen	Personen
UCI	Wielrennen	Organisaties
KNWU	Wielrennen	Organisaties
Tirreno-Adriatico	Wielrennen	Producten
Omloop Het Nieuwsblad	Wielrennen	Producten
Koppenberg	Wielrennen	Locaties
La Montagne de Lure	Wielrennen	Locaties
literature cycling	Wielrennen	Hogere concepten / onderwerpen
professional cycling developments	Wielrennen	Hogere concepten / onderwerpen
Maij-Weggen	Noord-Brabant	Personen
Vreeman burgemeester	Noord-Brabant	Personen
WonenBreburg	Noord-Brabant	Organisaties
Universiteit van Tilburg	Noord-Brabant	Organisaties
vergunning Brabant	Noord-Brabant	Producten
wandelroutes Brabant	Noord-Brabant	Producten
Westpoint Tilburg	Noord-Brabant	Locaties
Parade Den Bosch	Noord-Brabant	Locaties
Brabant crisis	Noord-Brabant	Hogere concepten / onderwerpen
Tilburg overval	Noord-Brabant	Hogere concepten / onderwerpen

APPENDIX B: DATA DEELEXPERIMENTEN

De data van de verschillende deelexperimenten zijn opgenomen in het Excelbestand **RM-zoekvragen-resultaten-analyse-11062009.xls**.

Dit Excelbestand beschikt over de volgende tabbladen:

- **Resultatenoverzicht**
  Bevat de belangrijkste resultaten van de deelexperimenten
- **Resultaten naar onderwerp**
  Bevat de geregistreerde data van de 50 zoekvragen en de berekeningen voor het totale aantal zoekvragen en voor de uitsplitsing naar onderwerpscategorie
- **Berekeningen sortering soort**
  Bevat de geregistreerde data van de 50 zoekvragen en de berekeningen voor het totale aantal zoekvragen en voor de uitsplitsing naar soort
- **Categorieënmatrix**
  Bevat de coderingen van de verschillende categorieën
- **Figuur precisie deelexp 1 en 3a**
  Bevat de figuur van precisie voor deelexperimenten 1 en 3a
- **Figuur precisie deelexp 2 en 3b**
  Bevat de figuur van precisie voor deelexperimenten 2 en 3b
- **Hulptabblad**
  Is een hulpmiddel om de resultaten voor de soortcategorieën te kunnen verkrijgen

LITERATUURLIJST

A. Agarwal 2009

Amit Agarwal. ‘Do You Really Need Google Blog Search Anymore?’. <http://www.labnol.org/internet/who-needs-google-blog-search/7956/> (12 maart 2009), geraadpleegd op 7 juni 2009.

N. Agarwal e.a. 2008

Nitin Agarwal. ‘Identifying the Influential Bloggers in a Community’. *Proceedings of the international conference on Web search and web data mining (WSDM’08)*, 2008, pp. 207-217.

N. Agarwal & Liu 2008

Nitin Agarwal, Huan Liu. ‘Blogosphere: research issues, tools, and applications’. *SIGKDD Explorations Newsletter*. Vol. 10, No. 1, 2008, pp. 18-31.

Arrington 2008

Michael Arrington. ‘Europe-Focused Blog Search Engine Twingly Goes Into Private Beta’. <http://www.techcrunch.com/2008/04/02/europe-focused-blog-search-engine-twingly-goes-into-private-beta/> (2 april 2008), geraadpleegd op 7 juni 2009.

Bade 2007

David Bade. ‘Relevance ranking is not relevance ranking or, when the user is not the user, the search results are not search results’. *Online Information Review*, Vol. 31, No. 6, 2007, pp. 831-844.

Bar-Ilan 2005

Judit Bar-Ilan. ‘Information hub blogs’. *Journal of Information Science*. Vol. 31, No. 4, 2005, pp. 297-307.

Bar-Ilan e.a. 2009

Judit Bar-Ilan, Kevin Keenoy, Mark Levene, Eti Yaari. ‘Presentation Bias Is Significant in Determining User Preference for Search Results — A User Study’. *Journal of the American Society for Information Science and Technology*. Vol. 60, No. 1, 2009, pp. 135-149.

Beall 2008

Jeffrey Beall. ‘The Weaknesses of Full-Text Searching’. *Journal of Academic Librarianship*. Vol. 34, No. 5, 2008, pp. 438–444.

Blood 2004
Rebecca Blood. ‘How Blogging Software reshapes the online community’. *Communications of the ACM*. Vol. 47, No. 12, 2004, pp. 53-55.

Boyd 2006
Danah Boyd. ‘A Blogger's Blog: Exploring the Definition of a Medium’. *Reconstruction*. Vol. 6, No. 4, 2006.

Broback 2008a
Steve Broback. ‘Google Blog Search: Sucky for Brand Monitoring’. <http://webcommunityforum.com/2008/07/google-blog-search-sucky-for-brand-monitoring/> (28 juli 2008), geraadpleegd op 7 juni 2009.

Broback 2008b
Steve Broback. ‘A Comparison of Blog Search Engines for Brand Monitoring: Part 1’. < http://webcommunityforum.com/2008/08/a-comparison-of-blog-search-engines-for-brand-monitoring-part-1/> (28 augustus 2008), geraadpleegd op 7 juni 2009.

Broder 2002
Andrei Broder. ‘A taxonomy of web search’. *SIGIR Forum*, Vol. 36, No. 2, 2002, pp. 3-10.

Buckland 1991
Michael K. Buckland. ‘Information as Thing’. *Journal of the American Society for Information Science*. Vol. 42, No. 5, 1991, pp. 351-360.

Buckland 1997
Michael K. Buckland. ‘What is a “Document”?’. *Journal of the American Society for Information Science*. Vol. 48, No. 9, 1997, pp. 804-809.

Chowdhury 2004
G.G. Chowdhury. *Introduction to Modern Information Retrieval*. London: Facet Publishing 2004.

Chopin 2008
Kimberly Chopin. ‘Finding communities: alternative viewpoints through weblogs and tagging’. *Journal of Documentation*. Vol. 64, No. 4, 2008, pp. 552-575.

Diana 2008
Rob Diana. ‘Blog Search May Suck, But What Do We Really Want?’. <http://www.louisgray.com/live/2009/01/blog-search-may-suck-but-what-do-we.html> (3 januari 2009), geraadpleegd op 7 juni 2009.

Gray 2009
Louis Gray. ‘Technorati's Revenge? The Site Is Beating Google for Blog Reactions’. <http://www.louisgray.com/live/2009/03/technoratis-revenge-site-is-beating.html> (5 maart 2009), geraadpleegd op 7 juni 2009.

Hearst e.a. 2008
Marti A. Hearst, Matthew Hurst, Susan T. Dumais. ‘What Should Blog Search Look Like?’ *Conference on Information and Knowledge Management (CIKM). Proceeding of the 2008 ACM workshop on Search in social media (SSM’08).* 2008, pp. 95-98.

Herring e.a. 2004
Susan C. Herring, Lois Ann Scheidt, Sabrina Bonus, Elijah Wright. ‘Bridging the Gap: A Genre Analysis of Weblogs’. *Proceedings of the 37th Hawaii International Conference on System Sciences (HICSS-37)*. 2004.

Hitwise 2009
Hitwise. *Google Received 72 Percent of U.S. Searches in January 2009 (press release).* < http://www.hitwise.com/press-center/hitwiseHS2004/google-searches-jan-09.php> (24 februari 2009), geraadpleegd op 7 juni 2009.

Hylton 2008
Jeremy Hylton. ‘Update on indexing blogrolls’ <http://groups.google.com/group/google-blog-search/browse_thread/thread/75267bf8c4766b0c> (19 december 2008), geraadpleegd op 7 juni 2009.

Hylton 2009
Jeremy Hylton. ‘Blogroll fix for link: queries’ <http://groups.google.com/group/google-blog-search/browse_thread/thread/790b97e23bdd3509/8c9712c958b3c79e> (13 april 2009), geraadpleegd op 7 juni 2009.

Jansen e.a. 2000
Bernard J. Jansen, Amanda Spink, Tefko Saracevic. ‘Real life, real users, and real needs: a study and analysis of user queries on the Web’. *Information Processing and Management*. Vol. 36, No. 2, 2000, pp. 207-227.

Jansen e.a. 2008
Bernard J. Jansen, Danielle L. Booth, Amanda Spink. ‘Determining the informational, navigational, and transactional intent of Web queries’. *Information Processing and Management*. Vol. 44, No. 3, 2008, pp. 1251–1266.

Kellar e.a. 2007
Melanie Kellar, Carolyn Watters, Michael Shepherd. ‘A Field Study Characterizing Web-Based Information-Seeking Tasks’. *Journal of the American Society for Information Science and Technology*. Vol. 58, No. 7, 2007, pp. 999-1018.

Kirkpatrick 2008
Marshall Kirkpatrick. ‘State of the Blogosphere 2008: Technorati Numbers Indicate Blogging Is Niche and Slowing’. <http://www.readwriteweb.com/archives/state_of_the_blogosphere_2008.php> (22 september 2008), geraadpleegd op 7 juni 2009.

Kirkpatrick 2009
Marshall Kirkpatrick. ‘The State of Blog Search, 2009’. <http://www.readwriteweb.com/archives/the_state_of_blog_search_engines.php> (29 januari 2009), geraadpleegd op 7 juni 2009.

Lewandowski 2008
Dirk Lewandowski. ‘The retrieval effectiveness of web search engines: considering results descriptions’. *Journal of Documentation*. Vol. 64, No. 6, 2008, pp. 915-937.

Lewandowski en Höchstötter 2008
Dirk Lewandowski and Nadine Höchstötter. ‘Web searching: a quality measurement perspective’. In: Amanda Spink and Michael Zimmer (eds), *Web Searching: Multidisciplinary Perspectives*. Berlijn: Springer, 2008, pp. 309–340.

Lomborg 2009
Stine Lomborg. ‘Navigating the blogosphere: Towards a genre–based typology of weblogs’. *First Monday*. Vol. 14, No. 5, 2009.

Mayer 2007
Marissa Mayer. ‘Universal search: The best answer is still the best answer.’ <http://googleblog.blogspot.com/2007/05/universal-search-best-answer-is-still.html>, (16 mei 2007), geraadpleegd op 7 juni 2009.

Meijer 2007
Reinier Meijer. *Evaluatie van zoekmachines. Verslag project module Information Retrieval (UvA)*, 2007.

Merriam-Webster Online Dictionary 2009
Merriam-Webster Online Dictionary. <http://www.merriam-webster.com/>, geraadpleegd op 7 juni 2009.

Miller & Shepherd 2004
Carolyn R. Miller & Dawn Shepherd, 'Blogging as social action: a genre analysis of the weblog'. <http://blog.lib.umn.edu/blogosphere/blogging_as_social_action_a_genre_analysis_of_the_weblog.html> (2004), geraadpleegd op 7 juni 2009.

Mishne 2007
Gilad Mishne. *Applied Text Analytics for Blogs*. diss. Amsterdam: UvA 2007. <http://dare.uva.nl/document/46517>, geraadpleegd op 7 juni 2009.

Mishne & De Rijke 2006
Gilad Mishne & Maarten de Rijke. 'A Study of Blog Search'. *Advances in Information Retrieval: Proceedings 28th European Conference on IR Research (ECIR 2006)*, 2006, pp. 289-301.

Morrison 2008
P. Jason Morrison. 'Tagging and searching: Search retrieval effectiveness of folksonomies on the World Wide Web'. *Information Processing and Management*. Vol. 44, 2008, pp. 1562–1579.

Nardi e.a. 2004
Bonnie A. Nardi, Diane J. Schiano, Michelle Gumbrecht, Luke Swartz. 'Why we blog'. *Communications of the ACM*. Vol. 47, No. 12, 2004. pp. 41-46.

Nowson 2006
Scott Nowson. *The Language of Weblogs: A study of genre and individual differences*. diss. Edinburgh 2006.

Owens 2008
Simon Owens. 'Can Technorati Beat Google at Blog Search?'. <http://www.pbs.org/mediashift/2008/12/can-technorati-beat-google-at-blog-search357.html> (22 december 2008), geraadpleegd op 7 juni 2009.

Oxford English Dictionary Online 2009
Oxford English Dictionary Online. < http://dictionary.oed.com/>, geraadpleegd op 7 juni 2009.

Rayward 1994
W. Boyd Rayward. 'Visions of Xanadu: Paul Otlet (1868-1944) and Hypertext'. *Journal of the American Society for Information Science*. Vol. 45, No. 4, 1994, pp. 235-250.

Rosenberg 2008
Scott Rosenberg. 'Google Blog Search loses its bearings'. < http://www.wordyard.com/2008/12/15/google-blog-search/> (15 december 2008), geraadpleegd op 7 juni 2009.

Rubel 2008
Steve Rubel. 'The Sorry State of Blog Search Engines'. <http://www.micropersuasion.com/2008/12/the-sorry-state.html> (17 december 2008), geraadpleegd op 7 juni 2009.

Seo en Croft 2008
Jangwon Seo, W. Bruce Croft. 'Blog Site Search Using Resource Selection'. *Proceeding of the 17th ACM conference on Information and knowledge management (CIKM'08)*, 2008, pp. 1053-1062.

Smarty 2008
Ann Smarty. 'Blog Search Engines: The Complete Overview'. <http://www.searchenginejournal.com/blog-search-engines-the-complete-overview/7856/> (24 oktober 2008), geraadpleegd op 7 juni 2009.

Sun e.a. 2008
Aixin Sun, Meishan Hu, Ee-Peng Lim. 'Searching Blogs and News: A Study on Popular Queries'. *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR'08)*, 2008, pp. 729-730.

Tague-Sutcliffe 1992
Jean Tague-Sutcliffe. 'The pragmatics of information retrieval experimentation, revisited'. *lnformation Processing & Management*. Vol. 28, No. 4, 1992, pp. 467-490.

Technorati 2008
Technorati. 'State of the Blogosphere 2008'. <http://www.technorati.com/blogging/state-of-the-blogosphere/> (september 2008), geraadpleegd op 7 juni 2009.

Technorati 2009
Technorati. ‘Mainstream Media, We Can Index Your Blogs Too’. <http://technorati.com/weblog/2009/03/478.html> (2 maart 2009), geraadpleegd op 7 juni 2009.

Thelwall & Hasler 2007
Mike Thelwall & Laura Hasler. ‘Blog search engines’. *Online Information Review*. Vol. 31, No. 4, 2007, pp. 467-479.

Twingly 2008
Twingly. ‘Twingly News: BlogRank and Top 100’. <http://blog.twingly.com/2008/12/16/twingly-news-blogrank-and-top-100/> (16 december 2008), geraadpleegd op 7 juni 2009.

Universal McCann 2008
Universal McCann. *Power to the people - social media tracker: Wave.3*. <http://www.universalmccann.com/Assets/wave_3_20080403093750.pdf> (maart 2008), geraadpleegd op 7 juni 2009.

Walker 2005
Jill Walker. Blog (definition). In Marie-Laure Ryan David Herman, Manfred Jahn (editor), *Routledge Encyclopedia of Narrative Theory*. Routledge 2005.

Wikipedia 2009
Wikipedia. <http://en.wikipedia.org/>, geraadpleegd op 7 juni 2009.

Wilde 2008
Erik Wilde. ‘Deconstructing blogs’. *Online Information Review*. Vol. 32, No. 3, 2008, pp. 401-414.

Wilson 1999
T.D. Wilson. ‘Models in information behaviour research’. *Journal of Documentation*. Vol. 55, No. 3, 1999, pp. 249-270.