14
Search Engines and Expertise about Global Issues: Well-defined Landscape or Undomesticated Wilderness?

J. Fry, S. Virkar, and R. Schroeder

Summary This chapter investigates the 'winner-takes-all' hypothesis in relation to how academic researchers access online sources and resources. Some have argued that the Web provides access to a wider range of sources of information than offline resources. Others, such as Hindman et al. (2003), have shown that access to online resources is highly concentrated, particularly because of how Internet search engines are designed. With researchers increasingly using the Web and Internet search engines to disseminate and locate information and expertise, the question of whether the use of online resources enhances or diminishes the range of available sources of expertise is bound to become more pressing. To address this question four globally relevant knowledge domains were investigated using large-scale link analysis and a series of semi-structured interviews with UK-based academic researchers.

We found there to be no uniform 'winner-takes-all' effect in the use of online resources. Instead, there were different types of information gatekeepers for the four domains we examined and for the types of resources and sources that are sought. Particular characteristics of a knowledge domain's information environment appear to determine whether Google and other Internet search engines function as a *facilitator* in accessing expertise or as an *influential gatekeeper*.

14.1 Introduction

It is widely believed that the rapid diffusion of the Internet and the Web has transformed knowledge and expertise by widening access and making information available globally. Whilst there has been an exponential increase in the production and use of networked digital resources, little is known about the reach and impact of this form of distributed knowledge. Some have argued that information technology could have a 'democratizing' impact on knowledge and information (Dahl 1989), others have argued the opposite: that in the online world these resources have in fact become concentrated in a 'winner-takes-all' effect (Hindman et al. 2003), due in part to the link-based indexing algorithms of search engines and how such tools are embedded in information seeking practices. There is thus a need to

A. Spink and M. Zimmer (eds.), *Web Search, Springer Series in Information Science and Knowledge Management 14.*

determine the extent to which the Internet is reshaping access to knowledge and resources world-wide (Dutton et al. 2003), particularly in science where the Internet is fast becoming the primary medium for communication and collaboration between scholars.

In this chapter, we address the issue of 'winner-takes-all' in relation to the use of online resources within four research domains: Climate Change, Internet and Society, HIV/AIDS, and Terrorism. Together these domains represent a broad mix of urgent global issues addressed by both natural and social sciences. As these topics are also arguably highly current and relevant on an international level, they provide a good case for examining whether access to scientific expertise is being reconfigured.

A popular approach for studying the dynamics of knowledge domains and the online presence of actors in those domains is Webmetric analysis (Park and Thelwall 2005). Hyperlink studies in social science research, generally referred to as Webmetric analysis, draws on techniques and frameworks from the information science field of bibliometrics. To this end, our research synthesised Webmetric data (detailed results of the Webmetric analysis are reported in Schroeder et al. 2005) with data gathered from an interview series with UK-based academic researchers. The following chapter focuses on the analysis of the interview data and the extent to which the "Google representation" of the information environment of each domain overlapped with respondents' mental models of the core institutions, people and resources in their domain. The aim of the interviews was to obtain a well-rounded understanding of how researchers use online resources, including how they combine online and offline sources of information, their use of search engines, and what kinds of sites they use most frequently.

14.2 Previous Research on How Expertise is Accessed on the Web

14.2.1 Studying the Scholarly Web

With the advent of the Internet and the Web, new online resources have become available and electronic media are becoming increasingly important channels for social interaction. In previous research, conflicting views have been argued about whether the shift to online resources democratizes or concentrates access.

A prominent argument that 'winners-take-all has been made by Barabási (2003), who has argued that power law distributions apply to online networks of hyperlinks. This concept is known in bibliometrics and the sociology of science as the 'Matthew effect' ('unto every that hath, shall be given') or cumulative advantage (Merton 1988). Pennock et al. (2002) have refined this idea by suggesting that while the winner-takes-all hypothesis may apply to the Web as a whole, the balance of competition varied by domain-specific types of pages and when distributions of

links are compared for the same type of pages, they exhibit a more uniform pattern of connectivity. For example, university homepages will exhibit a more uniform pattern of connectivity to other university homepages (Thelwall et al. 2005).

The results that Web search engines yield are only partly determined by 'real' links. Ensuring that certain links rank highly among search results has also become commercially competitive, with firms specializing in 'search engine optimization' (Van Couvering 2006). High-ranking search results are valuable because they may draw customers to a site and draw them to advertising links which feature on search engine results pages (so called 'sponsored links'). One might assume that academic knowledge domains are less likely to be influenced by commercial factors affecting search results, however, the four case-study domains varied in their degree of market penetration (Walsh and Bayma 1996) and there was some overlap between commercial and non-commercial producers in their Webspheres. A Websphere being, *"a collection of dynamically defined digital resources spanning multiple Websites deemed relevant or related to a central theme or object"* (Schneider and Foot 2002). For example, pharmaceutical companies were heavily represented in the Websphere of the HIV/AIDS domain, often sponsoring sites that appeared in the top ten results of a Google search. Moreover, not-for-profit research organizations may also be engaged in competition for prominence among search results, such competition may penetrate into how the four domains are represented on the Web by Internet search engines.

14.2.2 What We Know about Online Information Practices

Although this study focuses on how domain factors influence Web searching and access to online resources amongst academic researchers, it is important to contextualise these influences within a wider understanding of the search paths and strategies that non-domain experts develop when trying to locate information and how their online search for information intersects with their offline information practices.

User studies of information seeking have shown that the Web is now a primary source of information for many people, with over 80% of Web searchers using Internet search engines to locate information. This is especially important to take into account as individuals are increasingly turning to the Internet as their primary source of expertise in critical areas of everyday life, such as health (Johnson et al. 2006).

People's attention span is brief when finding information on the Web, with Web researchers spending between 5 and 120 minutes for individual sessions (Jansen and Spink 2006). In their comparison of nine studies of Web search based on Web transaction logs, Jansen and Spink (2006) found that the average search session length is fifteen minutes and that this has remained stable from 1997–2002. Single-term queries counted for between 20–30% of all queries with an increasing trend for shorter queries. General Internet users most frequently search for people, places

or things (41.5% in 2002). Most pertinently, the cross-study comparison by Jansen and Spink (2006) confirmed that the viewing of only one results page is increasing, with the percentage of searchers viewing only one results page increasing from 29% in 1997 to 73% in 2002.

There have been a number of studies that have focused specifically on users' experience with the Google Internet search engine. Granka et al. (2004), for example, used eye-tracking to study how users interact with the list of ranked results from Google. Analyzing all behaviour before a user clicks on the first link or exits the list, they found that the time spent viewing the URL abstract on the Google results page was distributed equally between the first and second ranked URLs. Users substantially more often click on the link ranked first. After the second link fixation time drops off significantly, especially after the first 5 or 6 results. This is partly because typically only the first 5 and 6 links are visible without scrolling and once a user starts scrolling then rank becomes less of an influence for attention (Granka et al. 2004).

Adams and Blandford (2005) found that academics and some clinicians preferred to use the Internet rather than specialist digital libraries for accessing information due to the lower barrier to entry in terms of ease of use. The experienced clinicians in Adams and Blandford's (2005) study reported that the Internet was an important tool for accessing authoritative and timely information sources. There was a concern, however, that new members of the domain would not be able to differentiate sufficiently between valid and non-valid sources identified through Internet search engines. In fact, Adams et al. (2005) found that the hierarchical structure of clinical settings meant that senior clinicians often acted as information gatekeepers for junior clinicians.

Johnson et al. (2006) studied of how individuals seek information about inherited cancers identified some common pathways in how people move between online and offline resources. Their findings reveal that of the seven one-step pathways where only one resource was consulted, 78 of respondents (12.1%) consulted the Internet only; of the two-step pathways 79 (12.3%) consulted the Internet then the library, 41 (6.4%) the doctor then the Internet, and 36 (5.6%) consulted the library then the Internet. There were only two common three-step pathways and these were; Internet then library then doctor 51(7.9%) of respondents, and Internet then doctor then library 38 (5.9%).

Naturalistic studies of relevance judgments have shown that relevance is shaped by the content of the user's information environment. This is particularly the case when exploring new domains as in the translation work and boundary crossing of interdisciplinary scholars (Palmer and Neumann 2002), novice domain inhabitants, or non-expert lay people. As the four domains examined here involve several disciplines, and given that people are still uncertain about what extent sources and resources are moving online, it is plausible to assume that Google will play a gatekeeper role depending on the way in which search is used and online resources are structured.

While studies of online search practice have examined query reformulation, multitasking and successive searches (Spink et al. 2001; Wang et al. 2003) they

have not studied when, how and why users shift their search to a different source. In other words we know little about users' persistence with, or 'loyalty' to, a particular resource or Web-based search tool. According to the Pew Internet and American Life Survey (2005) Internet users tend to settle quickly on a single search engine and then persist with it, rather than comparing results from different search systems.

The winner-takes-all effect is therefore likely to depend on the type of Internet user, so that while domain experts are inculcated in the significance criteria that should be applied in selecting an information resource or source, non-expert information seekers will have a higher degree of uncertainty (Whitley 2000) in judging relevance, validity, authority and differentiating between various sources. The hierarchical ordering of resources and sources could have winner-takes-all ramifications in how people make sense of information and incorporate it into their decision making.

Though individual practices should not be overlooked, it is also important not to trivialize the influence that structural considerations (Solomon 1999) have in shaping information environments and practices. This is particularly the case when studying how professionals and scientists seek information as aspects of information practice such as relevance, selection of resources, sense making and decision taking will be influenced by what is considered valid, pertinent and timely by the domain community.

14.2.3 Domain Factors

There has been a strong tradition of domain analysis in human information behaviour research. This approach treats domains as discourse communities or communities of practice, rather than focusing on users in a generalized and context independent manner (Hjørland and Albrechtsen 1995). In following this approach, consideration is given to cultural aspects of domains such as knowledge structures, language, patterns of communication and cooperation and the use of information systems (Palmer and Neumann 2002).

Scholars typically rely on a core set of resources in producing knowledge. Palmer and Neumann (2002) describe how scholars extend their intellectual province through information work. They also found that in interdisciplinary domains, there is a need for translation work and boundary crossing across information environments. They argue that the imprecise language used in the humanities and social sciences is especially poor for identifying topic-based conservations across domain boundaries. This leads to what they describe as "excavating", which is the tracing of intellectual paths through sources and resources. Humanities scholars tend to refer to their research approach as detective work and descriptions of their practices show that they do follow leads to great lengths, in terms of both time and space. This practice has two important outcomes: it creates a relatively unique path of information seeking for each project and it brings scholars in contact with diverse information resources and many forms of technology, from the antiquated to the state of the art.

Related to the notion of excavating resources across domain boundaries is the concept of 'scatter': the degree to which relevant material is either concentrated within core disciplinary resources or produced and found across diverse fields and resources (Fry and Talja 2004). For example, in their study of Faculty use of electronic resources Vakkari and Talja (2005) found that in medicine 52% of respondents used publications mainly from their own field, in engineering this was 40%, and in the social sciences only 21% mainly used publications in their own field. The concept of scatter, as identified by Mote (1962), has been linked to interdisciplinary penetration although is not exclusively a symptom of it. Scatter has so far mainly been used in relation to the concentration of journals in a domain. Scholars in low scatter fields are served by a small number of highly specialized journals, whereas in high scatter fields, relevant materials are scattered across several disciplines and published in a large number of different journals (Vakkari and Talja 2005). Scatter of literature across domain information environments also influences the nature of search. For example, in high-scatter multi-disciplinary fields, where concepts are often contested, search strategies are typically developed around particular *conversations* (Tuominen et al. 2003), rather than directed searching (Fry and Talja 2004).

14.3 Approach and Methods

In order to help determine whether the winner-takes-all hypothesis applied to patterns of access to information in the four domains the interview series was used to validate[152] a subset of the Webmetric data. This sub-set comprised the 'Google representation' of each domain e.g. the most prominently indexed institutions, organizations, people and resources. It was derived by identifying the top thirty sites retrieved from searches using the following keywords :

• Climate Change	– 'Climate change'
	– 'Global warming'
	– 'Ozone depletion'
• Internet and Society	– 'Internet and society'
	– 'Internet research'
	– 'Internet Studies'
• HIV/AIDS	– 'HIV/AIDS'
	– 'HIV Infection'
	– 'HIV prevention'
• Terrorism	– 'Terrorism'
	– 'Terrorist organisation'
	– 'Terrorist network'

[152] The implications of the time-lag between obtaining the results of the webmetric analysis and their validation through expert interviews must be given some thought. Close to 6 months elapsed between the webmetric research and the bulk of the interviews – a fairly significant duration given that the Web is growing and changing at a rapid place.

Each of the four case-study domains was international in scope, although some had more of a national orientation in terms of resources and audiences than others. They were by and large interdisciplinary in terms of their epistemic structures and had a policy-related orientation in their outcomes. For example, the *climate change* researchers came from environmental science, biodiversity and physics. Their research interests included energy in developing countries, global energy and forecasting. The field has a strong international orientation in terms of its research concerns, institutions, information sources and patterns of dissemination.

The *Internet and Society* researchers constituted an almost trans-disciplinary topic coming from diverse disciplines such as political science, sociology, science and technology studies, public policy, media and cultural studies, and psychology. Research interests included: public identity management; e-Health; e-Learning, and the use of technology in everyday life. This area tended to be less international than Climate Change given that many of the institutions being studied have a particular national role in governance such as identity cards and health provision.

The *HIV/AIDS* domain is less interdisciplinary than the other three domains, possibly due to the greater degree of professional control over the field (Becher and Trowler 2001; Whitley 2000). Researchers came from fields such as quantitative social science and nursing science, with research areas including sexual health and health policy.

Finally, the *Terrorism* researchers came from diverse disciplines such as religious studies, political science and international relations. Their research interests ranged from religious violence to international security. In terms of geographic orientation this domain can be described as 'Global' in scope e.g. that this domain has a global dimension when the research organization is, for example, taking a world-wide approach to the topic, and local when the focus is on a particular set of organizations or similar.

In total twenty researchers were interviewed from universities in the UK, five from each domain, with a wide range of experience in the field and a range of specializations within the topic. Interviewees were asked about their research background, key institutions, groups and people in their research networks, and the variety of online resources they used. Questions also focused on their online search strategies, such as the tools they used for finding information, the keywords they used and what kind of entities they tended to search for e.g. people, groups or institutions. The interviews were recorded, transcribed in full and analysed using the Nvivo software for qualitative data analysis.

14.4 Changing Work Practices

14.4.1 Increasing Use of Online Resources

The interview participants were unanimous in their use of the Internet and Web in finding key information related to their work. They responded that they use the Web *"all the time"*, *"all the time, for everything"*, and they also described it as a

"vital tool". There was, nevertheless, great variation in how they used online resources and sources. They used search engines not just to find published material about their topic, but also for locating grey literature, for scoping out a new topic and finding out about the research activities of individuals.

It was interesting to observe that discussions about what kind of information could be gathered about *other* researchers or research groupings in their domains were also reflected in researchers' perceptions about their *own* Web presence. One of the Climate Change researchers, for example, noted the large amount of work required for 'showcasing' his group's research online and that generating visibility by posting on message boards also required effort and constant maintenance. This was also corroborated by two of the Internet and society researchers who agreed that *"it has really become quite a task, a chore that everybody hates"* (IS03) to create and maintain one's Website. Further, the variable amount of information about researchers or research groups also limited the helpfulness of online resources, with some people and groups having much more information available than others. Consistency, in terms of what types of information and material was made available on individual and group Web pages varied by domain. For example, Climate Change researchers in particular commented on such variability, whereas the Internet and Society researchers took it for granted that they could go to personal home pages to download articles.

14.4.2 Decreasing Use of Libraries

With one exception (an Internet and society researcher who said he still uses libraries as he always has done), researchers registered a decline in the use of libraries. This ranged from those who now almost never use libraries:

> I don't use any offline material … so if I can't get a journal online I don't use it … I hardly ever use books; I don't have the time now. (IS 03)

To others who simply noted that more material is available online now "*I find myself going less and less to the library … it has really changed my way of doing research*" (IS 02). Researchers appeared embarrassed to admit that they very rarely go to a physical library anymore.

14.4.3 Combining Online and Offline Resources in Various Ways

There are a variety of ways in which researchers combined online and offline resources. This varied according to the stage they were at in their search for relevant material as well as according to the currency of available information sources. This variation, however, needs to be put into the context that domains varied in the extent to which all of the relevant material is available online (Törmä and Vakkari 2004).

One of the Internet and Society researchers who was intellectually closer to the computer science, rather than the social science, aspect of the domain said that *'the ACM portal … contains almost everything in computing'*, so going to individual journals is no longer necessary'. In this case we see that some domains have highly centralized gateways to information. Even in domains where key sources may not be centrally organized under a single online resource researchers still tend to start their search online, "I used to go from offline to online … and now I go online first." (TM02).

A further point made by one of the Internet and Society researchers is the problem of publication lag in their domain, whereby printed sources come too late to be useful:

> … nobody really reads the [print] papers and the journals anymore at all. If you get it in a journal, you've waited too late (IS 01).

Despite the importance of the availability of online articles in this researcher's topic area, rather than creating a personal collection of digital sources on his local computer this researcher used Google as a tool for re-finding resources that he uses regularly:

> I have to use the Web for primary and secondary sources because I lose them all! It's quicker to find them again than to store them that's the amazing fact. (IS 01).

There may also be differences in what type of online material is sought depending on the task in hand. For one Internet and Society researcher policy documents needed to simply to be *"tracked down online"*, as a pre-defined source for the project, whereas for another project, it was a case of *"looking online for things to build up material"* in the first place.

The researchers also needed different materials at different times. For example, one of the Terrorist researchers sometimes locates journal articles and books online (secondary materials), and at other times is mainly looking for speeches (primary material). Within the domain of Terrorism research, differences in the type of material sought may be determined by whether the topic is current or historical. For example, legal cases in Terrorism will not be available *"until the draft has been approved and becomes law"*, therefore printed papers and books are necessary. This is also the case for historians of Terrorism, whereas *"for those who are studying current trends of movements…current responses and reactions by government … the [Internet] is an absolutely vital source"* (TM03).

14.5 Validation of Cybermetric Results

Despite the communitarian view often held of scholarly communities, a large body of research that focuses on the dynamics of scholarly communication and collaboration has found that there is a strong winner-take-all or cumulative advantage effect in science, whereby over time researchers with an initial advantage in a domain obtain even greater advantage in the reputation of their research and control

over that domain. The same has arguably happened with regard to the Web presence of individuals, groups, organizations and institutions, with some Websites becoming increasingly central and dominant as information resources. Online hierarchies, however, do not necessarily represent offline status (Caldas 2005).

Findings from our Webmetric results indicated that a small number of cliques comprising of the most highly-linked sites existed within each of the four domains. These cliques were located at the top of a steep curve of the most highly linked sites (Schroeder et al. 2005), demonstrating a power-law distribution or *power law* tail for each of the domains. The presence of such a distribution mathematically indicated the 'winner-take-all' hypothesis, and implied that for each topic area, some sites are exponentially better connected within the network, with only a small share of Web nodes receiving or providing many links while the bulk of the nodes have only a few in-links or out-links each.

A limitation of large-scale Webmetric analysis, however, is that the social and institutional phenomena underlying hyperlink patterns are difficult to interpret (Thelwall 2006). It was necessary, therefore, to validate the Webmetric findings with active researchers in each of the domains by asking them about their information practices and characteristics of the information environment at the domain level. In addition, participants were presented with the 'Google representation' derived from the keywords listed in Sect. 3. Respondents were asked how well the Google representation mapped onto their own mental model of the domain – their individual perception of what constitutes the core set of resources and sources (It is important to note that this is different to a mental model they may have at any one time in relation to a situational information need).

Any overlaps or inconsistencies between the Google representation of each domain and the participants' own mental model was further validated by coding the Websites, institutions, organizations, people and other resources they reported using throughout the interview transcripts and then comparing this list with their responses to the Google representation. For example, a comparison was made between how the URLs within the Google representation were distributed across top level domain and top level country code domain names, and the institutional and geographical dimensions of the self-reported model of participants' information environments. The top ten URLs in each of the Google representations tended to be dominated by the large US-based Top-level country code domain names: .org, .com, .gov, and .edu. The break-down of top level domain names for the top 30 Google results across each of the four cases are shown in Table 14.1:

In general, participants recognised, but tolerated, the U.S. bias in the Google representation. Not only was this bias accepted, but it was also anticipated based on their experiences of using Internet search engines to locate sources and resources. To counter this bias some participants reported tailoring their searches or made use of Google's country specific indexes, depending on the geographic orientation of their research.

We found that the extent to which the two domain views, Webmetric versus inhabitants, mapped onto one another was determined by the geographic orientation of the domain, characteristics in networks of excellence and individual perceptions of Google's effectiveness.

Table 14.1 Break down of Google representation by top-level country code domain name

	Top-level country code domain name				
	.org	.com	.gov	.edu	Other
Climate change	13	3	5	1	.co (1) .ca (1) .ac (1) .ch (1) .int (2) .net (2)
Internet and society	6	10	0	14	
HIV/AIDS	12	4	7	2	.ca (3) .int (2)
Terrorism	12	5	5	3	.mil (2) .net (1) .gov.uk (1) tr (1)

14.5.1 Geographic Orientation of Field

The interviews revealed that there was only a limited overlap between the Google representation of each of the case study domains and the researchers' mental models of key networks, structures and organizations. Researchers reported that many of the key online resources in their domain were missing from the Google representation. The extent of the overlap appears to be domain dependent, with those researchers working within a more nationally orientated information environment reporting less of an overlap. For example, the HIV/AIDS researchers reported using national sources and resources, such as the *British Journal of Sexual Health*, UK-based charity organizations, such as the Terrence Higgins Trust, and national statistics, such as those distributed by the Office of National Statistics, and public sector organizations, such as the Health Protection Agency, but none of these appear in the top thirty Google results for generic domain keywords (even when the search was repeated using Google.co.uk). Climate Change researchers, on the other hand, for whom the geographical boundaries of research were far more 'international', were able to recognise many more Google results on the Climate Change validation sheet.

14.5.2 Networks of Excellence

In addition to the gaps that participants identified in response to the direct validation of the Webmetric data there was also a discrepancy between the organizations, institutions, people and resources that they reported using during the course of the interview and the Google representation. This was particularly true for the Web pages of academics and academic institutions that the respondents frequently used.

In the cases where participants recognised some of the top sites from the list, or named key institutions, groups or people that did appear in the top thirty results, those identified were unlikely to appear in the top 10 results. The low ranking of some of the institutions, resources and people they perceived to be core in their domain's information environment surprised some participants, such as this HIV/AIDs researcher, *"I'm surprised that the W.H.O. [World Health Organization] doesn't figure higher!"* (HA03).

One of the Terrorism researchers, on the other hand, accepted the low ranking of one of the top resources in their domain:

> … you've got the M.I.P.T. [Memorial Institute for the Prevention of Terrorism] database on here – absolutely crucial – halfway down the left column … The M.I.P.T. database is really outstanding! (TM 02).

The M.I.P.T database is branded by its producers as the 'Terrorism Knowledge Base' and is a non-profit organization with a remit to prevent terrorism in the U.S. and provides access to statistics about global terrorist incidents. It is interesting to note that if a Google search is run using the keyword 'Terrorism', the M.I.P.T. database appears as the 11th result, and typically this would be the first link on the second page of results. Such a difference between perception and actual Web-presence could be crucial in terms of online visibility.

Respondents from all four domains reported that they did not go beyond the first page or first ten links of Google results. The number of links that respondents typically viewed was dependent upon whether or not they were multi-tasking and how much time they had available:

> It depends if I've got a lot on ice and depends on time. Say I know I've got to search for a lot of things and I've got this session, I might make a decision after the first 10 [results] to stop. If I've got a bit more time, then that's 20. If I've got a lot more time I go to the first 30. I won't go beyond the first 30. And if you ask out of those three what happens the most, I'd probably say 10 because I'm always busy. (IS05)

Persistence with a particular set of results also depended on whether the researchers believed it was worth sifting through a large volume of irrelevant material to unearth 'gems'. The Terrorism researchers were more likely to persist with a particular set of search results, 'excavating' links in a similar way to the interdisciplinary humanities scholars observed by Palmer and Neumann (2002):

> … there's a great deal of rather boring work in culling and identifying sources, but you can't afford to not do it in case you miss some outstanding new input, and this is rather expensive in time and the resources of researchers, but there's no alternative to actually looking hard through what's available on the Net. (TM03)

This search behaviour was different to that described by the Internet and Society researchers who described more directed searching:

> [I search] for very specific things. I try to be as specific as possible otherwise you get too much nonsense. (IS03)

Fry and Talja (2007) have linked directed searching to scatter of relevant material across domain boundaries and the comparative findings reported here seem to corroborate their argument.

14.5.3 Preconceptions of Google Effectiveness and Tolerating Irrelevance in Search Results

Despite recognition by respondents that Google is a blunt instrument in terms of seeking information, and UK-based researchers' recognition of a persistent US bias in its indexes, it was still the main tool for finding sources and resources on the Web. As Fallows (2005) has argued, trust in Internet search engines amongst Internet users tends to be high and users often persist with a particular engine. When asked what alternate strategies they used if Google failed to retrieve relevant results, most respondents preferred to change the keywords or phrases they had used for a search and persist with Google, rather than change search engine.

> No, not another search engine - I stick with Google, but what I sometimes do is change the keywords. For example, what I typed in second I put in first place, which may also make a difference. Also I use other keywords, or when I get too many hits specify more and use a third or fourth keyword (IS02)

There was a general perception that the source they were looking for was available through Google if only they could 'hack' the indexes in the correct way:

> I'll start again, but as they say 'modify your search'… I assume that I can actually get it out of Google if only I've got the wit to get the search right. (IS01)

Respondents generally had preconceived notions about the effectiveness of Google. For example, they were aware that not all the results obtained in a Google search were going to be relevant. In fact, all of them expected the search engine to come up with what one researcher termed 'slash-and-burn kinds of pages' (TM02). Despite this, however, they preferred Google for its clean interface and perceived ease of use, and appeared convinced that Google could give them the results they were looking for should they only persist and tailor their searches through altering key words and narrowing down search terms.

Persistence with a particular search was also largely influenced by the searcher's perception of the overall quality of result set retrieved. Respondents reported assessing relevance and validity on two intellectual levels simultaneously. On one level, the respondents' approach to determining the quality and relevance of a hit was purely 'rational' (Pharo and Jarvelin 2006), and involved skimming through the URL abstract provided by Google. On another level, the respondents' approach to determining relevance was 'heuristic', bringing into play the subject-knowledge and particular experience of the researcher in judging the 'respectability' of the source or organization hosting the page and in evaluating site-content, with a clear preference for information coming from a 'reputable' person or institution. This held true across the four domains both for junior and senior researchers:

> To be honest [assessing the quality of search results] is almost subliminal – looking for words and sites that you think would be respectable. Normally I would skim through the words that come up and then I would look at the Web address. If it was some kind of non-entity of a Non Governmental Organization, then I might ignore it, if it was the World Bank or the United Nations I might have a look. (CC01)

Ensuring the validity of primary data, e.g. statistics, was a particularly important concern for Terrorism researchers who felt that, owing to the highly sensitive nature of the issues they dealt with as well as the difficulties they faced in identifying 'legitimate' sources of data, any information collected from a search should be subjected to a stringent quality control process.

> … [the problem is] most people don't sufficiently vet what they're looking at. I'm very, very careful about it, especially in a subject like mine where everything depends so upon subjectivity and perception, at least to some degree. (TM02)

In fact, the Terrorism researchers stressed the importance of following a two stage quality control process: in the first instance using their own judgment to determine what they felt was valid from within a list of search results, and then cross-checking the accuracy of the data by corroborating it with other sources of information, particularly with experienced colleagues in the field:

> … corroboration is just as important when you're using a source off the Net as it is when you're using traditional media. If there's only one report on one particular site which said X did Y or is responsible, claimed responsibility, for doing this on a certain date, and you can't find anybody else who said this, even [if it's] in one of the most respected sources within that country, then you begin to worry that they've made a mistake or that somebody's trying to put false information into their account. You know…corroboration is extremely important in our field. (TM03)

It is important to bear in mind that the disparity between respondents' positive response to the Google representation of the domain-level information environment and what was identified as missing according to their reported use of online resources may have been due to the fact that they situated themselves at an early stage of information seeking, a stage at which Kuhlthau (1993) argues users are likely to hold a rather general standard of inclusion. Users tend to hold different relevance perceptions at different stages of information seeking. During the early stages of search formulation users tends to be more receptive to topically relevant items presented to them, whereas in the later stages following query formulation, the user tends to be more discriminating in identifying items only pertinent to their personal information need (Kuhlthau 1993).

14.6 Web-based Search Strategies

Though there were similarities in Web-search strategies across each of the four case study domains, there were also important differences. For example, while respondents reported using Google almost to the exclusion of all other generalist Internet search engines, the role that it played in their wider information environments varied considerably. In the HIV/AIDS and Internet and Society domains, for instance, Google is mainly used as what Beauvisage (2004) calls an "aide memoir", a locating tool for known sources. As one Internet and Society researcher noted:

> Very rarely would I put in a general query through Google – I think it [any query I do put in] would normally lead specifically to a policy article, newspaper or another article. I'm normally following up specific leads. (IS04)

The HIV/AIDS researchers, in particular, described quite distinct ways by which they found information and literature online, the predominant mode being Weblinks embedded in the body of email correspondence. These researchers tended to go to aggregated literature databases such as *PubMed* in the first instance, and then transferred their search to Google if they could not find the material they wanted:

> … if I had heard of a paper and could not access it through something like *PubMed* then I would do a Google search and try and find a Webpage for that author and see if they had a link to their paper. (HA01)

In response to the question whether they more often searched for people, topics or institutions, almost all researchers responded that they more often than not searched for a topic, looking for a specific information source. Occasionally, researchers looked for the home pages of people they had either met at conferences or whose work they were familiar with and wished to consult. This suggests that Web-based information seeking within the HIV/AIDS domain is driven by looking for known sources, rather than searching the Internet in an exploratory way.

In contrast, for researchers of Terrorism, Google plays a more central role in exploring the object of research and identifying relevant sources. This may be due to the amorphous, shadowy nature of the subject matter itself – Websites of terrorist groups and the message-boards, chatrooms and blogs associated with them are constantly being shut down by national intelligence agencies, only to resurface with new Web-addresses, and the only way to locate these and other sources like them is for researchers to 'excavate' resources across a range of resources and domain boundaries.

This varying role of Google was not solely contingent on domain, however, but also varied within domain according to what stage individuals were at in relation to a particular task. For example, while there was a core set of known Web-based resources within the HIV/AIDS domain, the following researcher also noted using Google as an exploratory tool:

I think the thing is, when I am doing just general background research I tend to use Google more. If I am writing an article and I need to find specific information or in my mind I know that there's something out there on this topic that I haven't quite managed to track down, then I'll transfer to more specific journal site searches. (HA03)

The Climate Change researchers, used search engines both for exploratory searches as well as searching for specific information or datasets or literature on people's home pages. Like HIV/AIDS researchers, researchers of Climate Change mentioned sharing papers or links to papers via email as a common way of obtaining information:

> … I think they [other Climate Change researchers] are really good at sharing papers over email – if you ask for a paper you normally get it pretty promptly, or a link to it. (CC04)

One possible explanation for differential domain patterns in the role of Google and other Internet search engines as information seeking tools could be the extent to which important documents are scattered across domain boundaries (Bates 1996). The consequence of this for Web searching is that in low scatter

fields, resources and sources can be found using a clearly circumscribed set of keywords and are likely to be produced by a limited number of dominant gatekeepers. Of the case studies, HIV/AIDS was the domain with the least scatter and this could explain why Google was used more as an 'aide memoir' than as an exploratory tool. Terrorism and Climate Change researchers on the other hand described their domains as scattered in terms of resources and respondents reported using Google for finding diverse sources more than in the other two case studies:

> Well, I've mentioned the World Bank, it's very important, but apart from that it's essentially very scattered. There are individual pages of bilateral and multilateral donors – the United Nations obviously, the International Energy Agency and specialized groups working on everything from renewable energy to national policies on cooking. (CC01)

Scatter also influences the degree to which directed searching, chaining or browsing will the most rewarding search technique (Fry and Talja 2007).

14.7 The Role of Gatekeepers

The characteristics and role of the predominant gatekeepers varied across each of the four domains. In this section we describe these differences and discuss their implications for degrees of 'winner-takes-all' on the Web.

The interview responses indicate a differentiated shift towards the decentralization of gatekeepers on the Internet. For example, in Climate Change 'hybrid research centres' produce and disseminate important sources; and policy or academic research centres are key producers of information sources in Internet and Society research. Although not-for-profit organizations were key producers and disseminators of information and played an important gatekeeping role in the HIV/AIDS information environment, traditional gatekeepers such as publishers still maintain a central position in this domain because of the continued importance of peer-reviewed articles disseminated through discipline-centric aggregated databases such as *PubMed Central*.

The information environment of the Terrorism researchers was similar to that of the HIV/AIDS researchers in that, while non-governmental and not-for-profit organizations play a central role in disseminating primary information resources, publishers still had an enduring role as gatekeepers to academic research. In Terrorism, dissemination of research via books plays a major role in the scholarly communication system and still remains closely interrelated to the recognition and reward system. Research in Terrorism is of a sensitive nature, which may account to some extent for the sustained importance of the traditional gatekeepers such as publishers.

In contrast, the gatekeepers in the information environments of the Climate Change and the Internet and Society researchers were more decentralized. This meant that rather than access to information being coordinated by a predominant gatekeeper there were multiple gatekeepers providing specific resources in niche areas.

This variation in the characteristics and role of gatekeepers in the information environments of each of the four case study domains appears to be influenced by a number of domain-specific intellectual and social factors. This includes the types of data used in each domain, which, except for the *Terrorism* researchers, was primarily quantitatively oriented. For example, the Climate Change participants relied heavily on international and national statistics produced by The World Bank and the International Energy Agency; the Internet and Society researchers frequently used national statistics based on Internet surveys produced by academic research centres; the HIV/AIDS researchers cited national health statistics, such as those produced by the Health Protection Agency in the UK, as leading sources of information. For the Terrorism researchers, on the other hand, news sources and public speeches were a source of primary information, but there was also a heavy reliance on secondary sources such as academic publications.

There was also variation at the domain level in the extent to which researchers had developed practices to by-pass gatekeepers such as publishers and libraries. Amongst Internet and Society researchers, this was making full-text articles available on academic home pages:

> … it's getting more and more important to have a good homepage and I really like the way that people publish all their papers and so forth on the Internet. It's so helpful to go to somebody's home page and to know that everything is there. (IS02)

Whereas in the information environments of Climate Change, Terrorism and HIV/AIDS research, academic homepages have a much lower information valency. This variation may also be explained by the nature of each domain's Websphere (Fry 2006), in terms of the extent to which it is academically oriented or oriented towards not-for-profit organizations. Table 14.1 shows the top-level domains of the URLS in the Google representation for each domain. The representation of the academic domain (e.g. .edu and .ac.uk) is low within each of the case studies except for Internet and Society, whereas the .org domain accounts for the largest percentage of domains across HIV/AIDS, Climate Change and Terrorism.

The Climate Change and Internet and Society respondents reported mainly using the Web for finding policy documentation and survey reports. Thus, there was no alternate way of accessing this information other than the not-for-profit organizations with a national or international remit for producing such knowledge and data. Within Climate Change attempts had been made to centralize disparate sources across multiple organizations and institutions, but there had been difficulties in centralizing, integrating and maintaining scattered resources:

> There have been a number of attempts in the past to try and collate this, through a meta-site, but normally they are not successful because people don't put in enough effort to keep them updated. (CC01)

Interestingly, although publishers play a key dissemination and access role in HIV/AIDS and Terrorism, they are absent from the Google representation, which may reflect their low visibility in the wider domain Websphere.

14.8 Implications for Web Search

Where the information environment of a domain is highly-structured, well-organized and dominated by a limited number of gatekeepers search and other forms of information seeking are likely to also be structured, highly focused and predictable, as with the HIV/AIDS domain where researchers go to *PubMed Central* in the first instance using a well-defined set of keywords. In these cases, therefore, search is directed within a 'well-defined landscape'.

This contrasts with the information environments of the Terrorism and Internet and Society domains where resources are scattered across a diverse range of gatekeepers and domains. Consequently, concepts are often contested, which leads to more open-ended undirected searches and increased uncertainty with regard to the appropriate keywords to search. Rather than search being for a particular specialized concept it is often for individual researchers, institutions or general concepts. In these domains access to online resources is more likely to depend on the indexing algorithms of Internet search engines and the online presence of particular institutions, organizations, people and resources. We describe this type of information environment as 'undomesticated wilderness'. In short, the Websphere of a domain can be seen as a realm that in some domains is centrally structured around traditional gatekeepers, and in others more decentralized and fragmented across traditional and emergent gatekeepers. Access to sources of expertise in the online realm will be strongly shaped by this organization or lack of it.

For example, traditional gatekeepers with a high-degree of offline status, such as the major journal publishers, also contribute to this degree of organized-ness of the Websphere and shape the extent of Google's impact on 'winner takes all'. If, therefore, there are predominant centralized knowledge gateways within a domain's information environment, such as *PubMed Central* within HIV/AIDS, then Google mainly functions as a *facilitator* in accessing them. If, on the other hand, a domain's information environment is decentralized (Fry and Talja 2007) and the production of resources is fragmented across a range of different types of institutions and organizations e.g. academic, not-for-profit or commercial, Google becomes an *influential gatekeeper*. In terms of understanding the potential 'winner-take-all' effect on the Web it is therefore useful to identify more specifically *which* resources are concentrated online as well how new sources of information displace and complement those that have been traditionally used.

A different example, perhaps not of 'concentration' but rather of how access to expertise is shaped, is the degree to which fields are oriented to a national scientific, practitioner or lay-audience. The more nationally oriented a resource or producer is, the more likely it is to be marginalised by the current Internet search engines in popular use. This is particularly apparent when contrasting health-related topics, such as HIV/AIDS, as against more globally oriented topics such as Climate Change and Terrorism, though the U.S. bias of search results cuts across all four topics. Even if this bias does not relate directly to the extent of the well-organizedness of the Websphere it is nevertheless closely connected because 'organizedness' can be related to the degree of 'boundedness' of the information landscape.

14.9 Conclusions

Our qualitative interview findings corroborate the quantitative Webmetric results (Schroeder et al. 2005) that there is no uniform 'Winner-takes-all' effect in the use of online resources. Instead, there are different kinds of gatekeepers for the four topics we examined and for the types of information that are sought. We found the effect to be differentiated according to four factors: geographic orientation of knowledge domains; strength or weakness of networks of excellence; the scatter of material across disciplinary boundaries and the role of traditional gatekeepers. It is therefore important not just to identify a concentration or democratization effect, but rather to refine under what circumstances the search for expertise will be dominated by certain results and exhibit biases, and when, instead, researchers will be led to the resources they seek and to a variety of results. Particular characteristics of a domain's information environment will determine whether Google and other Internet search engines function as a *facilitator* or as an *influential gatekeeper*.

Web search engines, and Google in particular, thus exercise a gatekeeping function, at the same time that they enable researchers to find their way in the new online environment. But this environment can be a well-organized landscape or a less-well charted wilderness. As the production and use of online resources continues to grow, it will become increasingly important whether search can find its way through these different types of landscapes. For topics such as those examined here, which in some way cross the boundaries of established disciplines, there is the additional question of the extent to which online resources will transcend or reconfigure established bounds of expertise. Such a shift will necessitate libraries and publishers – not to speak of researchers and institutions that produce and use material online - to realign their strategies for organizing services and content accordingly.

References

Adams A, Blandford A (2005) Digital libraries' support for the user's information journey. Proceedings of the 5th ACM/IEEE-CS joint conference on digital libraries 2005, Denver, CO, USA, June 7–11, pp 160–169

Adams A, Blandford A, Lunt P (2005) Social empowerment and exclusion: a case study on digital libraries. ACM Transactions on Computer–Human Interaction, 12.

Barabasi A (2003) Linked: the new science of networks. Perseus Books, New York

Bates MJ (1996) Learning about the information seeking of interdisciplinary scholars and students. Library Trends 45: 155–164

Beauvisage T (2004) A semantics of users' paths through the web. Unpublished PhD, University of Paris X: Nanterre, Paris (translated by Van Couvering, Web Behaviour: Search engines in context, draft paper)

Becher T, Trowler P (2001) Academic tribes and territories: intellectual inquiry and the culture of disciplines (2nd ed.). Open University Press, Milton Keynes

Caldas A (2005) On the origins of the web species and complexity. Paper presented to the New Approaches to Research on the Social Implications of Emerging Technologies Workshop, Oxford, 15–16 April

Van Couvering E (2006) Web behaviour: search engines in context. Available at: http://personal.lse.ac.uk/VANCOUVE/. Accessed 3 August 2006

Dahl R (1989) Democracy and its critics. New Haven, CT: Yale University Press.

Dutton WH, Gillet SE, McKnight LW, Peltu M (2003) Broadband Internet: the power to reconfigure access. Forum discussion paper no. 1, Oxford Internet Institute, August 2003

Fallows D (2005) Search engine users: Internet searchers are confident, satisfied and trusting – but they are also unaware and naïve. Report of the Pew Internet and American Life Project. Available at: http://www.pewInternet.org/pdfs/PIP_Searchengine_users.pdf. Accessed 3 August 2006

Fry J (2006) Studying the scholarly web: how disciplinary culture shapes online representations. Cybermetrics, 10. Available at: http://www.cindoc.csic.es/cybermetrics/vol10iss1.html. Accessed 3 August 2006

Fry J, Talja S (2004) The cultural shaping of scholarly communication: explaining e-journal use within and across academic fields. In: ASIST 2004: Proceedings of the 67th ASIST Annual Meeting 41. Medford, NJ.: Information Today

Fry J, Talja S (2007) The intellectual and social organization of academic fields and the shaping of digital resources. Journal of Information Science 33: 115–137

Granka LA, Joachims T, Gay G (2004) Eye-tracking analysis of user behavior in www-search. Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp 478–479

Hindman M, Tsioutsiouliklis K, Johnson J (2003) Googlearchy: how a few heavily-linked sites dominate politics on the web. Available at: http://www.princeton.edu/~mhindman/googlearchy–hindman.pdf. Accessed 7 February 2006

Hjørland B, Albrechtsen H (1995) Toward a new horizon in information science : domain-analysis. Journal of the American Society for Information Science 46: 400–425

Jansen BJ, Spink A (2006) How are we searching the world wide web? a comparison of nine search engine transaction logs. Information Processing and Management 42: 248–263

Johnson DE, Case DO, Andres J, Allard SL, Johnson NE (2006) Fields and pathways: contrasting or complementary views of information seeking. Information Processing and Management 42: 569–582

Kuhlthau CC (1993) Seeking meaning: a process approach to library and information services. Ablex, Norwood, NJ

Merton RK (1988) The Matthew effect in science II, Isis 79: 606–623

Mote LJB (1962) Reasons for the variation of information needs of scientists. Journal of Documentation 18: 169–175

Palmer CL, Neumann LJ (2002) The information work of interdisciplinary humanities scholars: exploration and translation, Library Quarterly 72: 85–117.

Park HW, Thelwall M (2005) The network approach to web hyperlink research and its utility for science communication. In Christine Hine (Ed.) Virtual methods: issues in social research on the Internet. Berg, Oxford: 171–181

Pennock DM, Flake GW, Lawrence S, Glover EJ, Giles CL (2002) Winners don't take all: characterizing the competition for links on the web. Proceedings of the National Academy of Sciences 99: 5207–5211

Pew Internet and American Life Survey (2005) Search engine users: Internet searchers are confident, satisfied and trusting – but they are also unaware and naïve. Available at: http://www.pewInternet.org/PPF/r/146/report_display.asp. Accessed 12 July 2006

Pharo N, Jarvelin K (2006) Irrational searchers and IR-rational researchers. Journal of the American Society for Information Science and Technology 57: 222–232

Schneider SM, Foot KA (2002) Online structure for political action: exploring presidential websites from the 2000 American election, Javnost (The Public) 9: 43–60

Schroeder R, Caldas A, Mesch G, Dutton W (2005) The world wide web of science: reconfiguring access to information, First International Conference on e-Social Science, Manchester 22–24 June, Available at: http://www.oii.ox.ac.uk/research/project.cfm?id=22. Accessed 6 March 2007

Solomon P (1999) Information mosaics: patterns of action that structure. In Wilson, T., and Allen, D.K. (Ed.) Exploring the contexts of information behaviour (pp. 150–175). UK. London: Taylor Graham

Spink A, Wolfram D, Jansen BJ, Saracevic T (2001) Searching the web: the public and their queries. Journal of the American Society for Information Science and Technology 52: 226–234

Thelwall M (2006) Interpreting social science link analysis research: a theoretical framework. Journal of the American Society for Information Science and Technology 57: 60–68

Thelwall M, Vaughan L, Björneborn L (2005) Webometrics. Annual Review of Information Science and Technology 39: 81–135

Törmä S, Vakkari P (2004) Discipline, availability of electronic resources and the use of Finnish national electronic library – FinELib. Information Research 10 Available at http://informationr.net/ir/10-1/paper204.html. Accessed 13 July 2006

Tuominen K, Talja S, Savolainen R (2003) Multiperspective digital libraries: the implications of constructionism for the development of digital libraries. Journal of the American Society for Information Science and Technology 54: 561–569

Vakkari P, Talja S (2005) The influence of the scatter of literature on the use of electronic resources across disciplines: a case study of FinElib. In: A. Rauber et al. (Eds.) ECDL 2005, LNCS 3652, pp 207–217

Walsh JP, Bayma T (1996) Computer networks and scientific work. Social Studies of Science 26: 661–703

Wang P, Berry MW, Yang Y (2003) Mining longitudinal web queries: trends and patterns. Journal of the American Society for Information Science and Technology 54: 743–758

Whitley R (2000). (2nd ed.) The intellectual and social organization of the sciences, Oxford: Clarendon Press.