James Caufield

The Myth of Automated Meaning

Abstract:

Most discussions of search engines focus on technology or user experience. By contrast, this paper asks about those who produce the recommendations that search engines gather. How are these people and institutions affected when search engines incorporate their work into search results, but no credit is given? The paper argues that the lack of attribution encourages the myth of automated meaning, the false belief that computers and algorithms have created rather than simply gathered these recommendations. It further argues that by concealing the role of these producers, search engines undermine public support for the individuals and institutions that create trustworthy recommendations, especially libraries. Because search engines borrow so extensively from public institutions and the public at large, their ethical obligations are far greater than previously recognized. The paper concludes with some comparisons between the ethical practices of libraries and those of search engines.

Agenda

Author:

James Caufield:

> New Mexico State University, P.O. Box 30006, MSC 3475, Las Cruces, New Mexico, 88003-8006.
> ☎ 505-646-5551, ✉ Caufield@nmsu.edu.
> Relevant publications:
>> - "Where Did Google Get Its Value?" *portal: Libraries and the Academy*, vol. 5, no. 4, October 2005, pp. 555-572.

Introduction

The recent development of search engines has solved one of the most vexing problems of the early Web, namely how to find valuable materials amid the vast number of web pages. *"Until Google, search engines helped a bit, but only a bit. They told you what was popular but not necessarily what you wanted to know. But Google's inventors found a way to put the most useful information at the top of the pile"* (McRae, 2004). Google did this by counting links between web pages and by studying user behavior, thus gathering meaningful recommendations from the Web itself.

Yet the brilliance of this success might blind us to its darker implications. Studies of search engines usually focus on technology or user experience. By contrast, this paper asks about the people and institutions who produce the recommendations search engines gather. While search engines are in the business of borrowing and aggregating trustworthy, relevant, and credible recommendations, they give no credit to the sources that provide them. The thesis of this paper is that by concealing the role of these producers, search engines undermine public support for those who create this meaning, especially libraries.

A few notes about method: private search engine companies are highly secretive about their methods and sources, making direct investigation impossible. This makes the paper in part a speculative effort, and so it pursues a line of thinking as far as possible. It advances what might be called a worst-case scenario, not because I know this to be true in all its details, but rather because I think this question needs to be considered in terms of its furthest possible ramifications.

As for intellectual abbreviations, I will usually call the information that libraries and search engines provide *recommendations*, but sometimes I will use more general terms such as *meaning* or *value*.

Finally, I have treated the positive contributions of Google elsewhere (Caufield, 2005); this paper will have a critical focus.

Review of Some Arguments Concerning the Ethical Obligations of Search Engines

Many reasons have been given for holding search engines to high ethical standards. First, using search engines is a practical necessity, as they are the *de facto* gatekeepers to information on the Web (Reider 2005: 27; Hinman 2005: 21; Neuberger 2005: 5). Thus a search engine such as Google

> *… plays an important role in the information many people rely on for choices in their lives. Cumulatively, the quality of Google's ranking impacts millions of choices made daily* (Wiggins, 2003).

This responsibility is intensified in that these tools are often used to find vital information (e.g., medical; see Fallows, 2005: 1, 6), and users place inordinate trust in search engines (Fallows, 2005: 27) which, as Rieder (2005: 31) and others argue, should not be exploited.

Moreover, the space that is accessed –the Web– is not private property, and it fulfills many of the functions of traditional public spaces such as libraries, museums, educational institutions, and public fora (Nissenbaum and Introna, 2004: 19-22). Finally, as search engines become more central to our information infrastructure, they assume important social and political roles, both in terms of providing the information essential to democratic decision-making, and in assuring the openness necessary for a free society (Letwin, 2004; Hinman, 2005: 21-2, 25). To this list of reasons for holding search engines to high ethical standards, this paper will add one more.

The Myth of Automated Meaning and Search Engines' Ethical Obligations

The most common way to describe Google is in technological terms. Oddly enough, this description is sometimes anthropomorphic, as in the following passage:

> *…the business story of Google is less interesting than the technological one: if information is power, then Google has helped change the world. Knowledge is measura-*

bly easier to obtain. Google works. Google knows (Achenbach, 2004).

Anthropomorphism is one version of the myth that attributes meaning to the machine, but it also serves as a reminder: computers do not know, and assigning relevance is a subjective determination, as is assigning other values (Blanke, 2005: 35-37; Rieder, 2005: 29). Thus Google is a semantic aggregator (Brooks 2004), a machine that only gathers meaning and does not create it. If we forget this, we succumb to the myth of automated meaning.

The implications of this insight have not been sufficiently noted. Most arguments concerning the ethical obligations of search engines point to the vitally important services these companies *provide to* the public. But perhaps the most telling argument concerns what search engines *take from* the public. So long as it is believed that search engine technology creates meaning, it is logical to think that search engines own this information. The recognition that search engines gather meaning from public sources implies a different conclusion. That search engines receive such a massive benefit from the public implies a correspondingly large obligation to the public, an ethical relationship perhaps more akin to stewardship than to ownership.

The remainder of this paper will investigate how and why search engines draw upon the work of public sources, particularly libraries.

Search Engines Need Trustworthy and Unbiased Results

The myth of automated meaning encourages us to think about search engines in terms of technology, so that we tend to forget or de-emphasize their human and ethical aspects. But consider an extreme case, the search service Overture, a competitor to Google, originally founded as Goto.com. In its earliest incarnation all search results were advertisements, with the top slot being given to the top bidder. The private interests of advertisers and the search engine were paramount, and disinterested service to the public was of little concern. Users spurned it: *"Overture soon figured out … that users did not want a search engine that was only advertising"* (Hansell, 2004). While Overture continues to include paid links (advertisements) in its search

results, Google has never allowed this practice, so it is hardly surprising that Google results are more relevant and precise than those of Overture (as reported by Goh and Ang, 2003). Thus relevance seems to be a product of ethics as well as of technology.[1]

Although relevance is commonly discussed as if it were the only value by which search engines should be judged, it is not. An exclusive focus on the relevance of results is perhaps appropriate when institutional structures have already assured the trustworthiness of the materials to be searched. This is the case in libraries, where quality is rather rigorously controlled, but it cannot be assumed in the Internet environment, where dubious materials abound. For instance, if a search engine produced recommendations for cancer treatments solely on the basis of advertising revenues and without regard to credibility, this information could indeed be relevant but still worth little. In order for search engine results to be valuable they need to be not only relevant but trustworthy as well.

An all-advertising search engine will not draw users and so, like most other commercial web enterprises, search engines need what the industry calls *'real estate,'* some non-advertising content that will draw *'eyeballs'* to whom it can present advertisements. For some web sites this is bird watching, for others sports news, etc. For search engines it is relevant and trustworthy recommendations. It seems that without this more-disinterested content, search engines will not draw the users they require in order to charge for advertising.

Users' Response to Search Engines that Exploit Trust for Private Benefit

There is evidence that search engine users believe they are receiving disinterested content, though often they are not. Many search engines accept money for including or placing advertising in their results, a fact that many writers on the subject see as problematic (Neuberger, 2005: 10-11; Himma

---

[1] Certainly other factors (including technological ones) play a role, but it stands to reason that when a search engine produces results according to two criteria, relevance and advertising revenue, rather than the criterion of relevance alone, then the relevance of results will be compromised.

2004: 474; Moxley et al., 2004; Scott, 2003; Nissenbaum and Introna, 2004: 14; Larsen, 1999: 22-3). Users are generally unaware of this practice, but once informed they recognize it as an ethical issue. An ethnographic study conducted by Leslie Marable of Consumer WebWatch found that users are frequently unable to distinguish between links that are paid (advertisements) and those that are not (Marable, 2003, p. 15), and this is because searchers have little conception of how search engines work (Marable, 2003: 7, 39-40; see also Neuberger, 2005, p. 5, citing Princeton Survey Research Associates 2002: 17; Fallows, 2005: 27). As Nissenbaum and Introna have observed, *"[g]iven the vastness of the Web, the close guarding of algorithms, and the abstruseness of the technology to most users, it should come as no surprise that seekers are unfamiliar, even unaware, of the systematic mechanisms that drive search engines. Such awareness, we believe, would make a difference"* (Nissenbaum and Introna, 2004: 16).

The Marable study shows that awareness does indeed make a difference. When participants in the study were informed that some search engine results were paid advertisements, *"each participant expressed surprise after learning about the search engine marketing practice. Some had negative emotional reactions"* (Marable, 2003; 39), including anger, a sense of betrayal, being somewhat stunned, disappointment, and a feeling of helplessness. Two people explicitly named this as an ethical issue, questioning *"… the trustworthiness of advertisers that chose to participate in search engine marketing programs"* (Marable, 2003; 18). *"The participants unanimously believed paid search links were too tough to recognize on many sites, and the related disclosure information was clearly written for the advertiser, not the consumer"* (Marable, 2003: 23). A similar example is given by Nissenbaum and Introna: *"[w]hen customers learned that Amazon Books, for example, had been representing as 'friendly' recommendations what were in reality paid advertisements, customers responded with great ire and Amazon hastily retreated"* (Nissenbaum and Introna, 2004: 16). As Fallows summarizes the situation,

> *"[d]espite … the explosion of paid placement sales as a source of revenue for search engines, users remain largely unaware of these issues – even though they still hold strong views of what they consider to be ethical practices by search engines. …[M]any users would object to the practices of search engines they rely on, if they*

> *understood more clearly how those engines actually operate"* (Fallows, 2005: 17).

In short, users tend to trust that search engine results will not be biased by private interest, and tend to take offense when they are informed otherwise. As Marable found, the depth of this misplaced trust is indicated by a rather astonishing belief:

> *"Our findings show Web searchers in this ethnographic study chose links from the first page of results nearly 50% of the time because they trusted the search engine to present only the best or most accurate results first. This <u>trust</u> led them to believe it was <u>unnecessary</u> for them to review later results pages"* (Marable, 2003: 38).

Speculations as to Why Users Trust Search Engines

This paper has argued that the quality of information is at least as much a function of ethical decisions as it is of technology. But it is also true that certain technologies have come to be associated with certain ethical practices. There is a long association between the technology of printing and a high quality of information. Of course this is not directly due to the technology, but rather to the culture that has developed in the print environment. Generally, anything published in paper is relatively trustworthy, as the publication process involves many steps of editorial selection and review. Even ephemeral media such as newspapers are held to fairly high standards, and the academic process is especially rigorous, involving a system of critical review on many levels. Thus the printed format has long been associated with a high level of credibility and trustworthiness.

Documents available on the Internet bear a strong resemblance to printed documents. Yet these Internet documents are not as trustworthy and reliable as printed documents because they usually do not have the institutional supports that would vouchsafe their quality. The likely scenario is that because of the similarity of formats users mistakenly extend their trust of printed materials to documents found on the Internet. Since this unwarranted trust of things on the Internet seems to have its origin in the traditions of the print culture, it could be said that digital illiteracy (at least in the sense of an overly-credulous attitude toward information on the Internet) has been caused by … literacy!

The misplaced trust in search engines is a subset of the misplaced trust of things on the Internet. The library online catalog is eminently trustworthy, at least in the sense of being free of commercial bias.[2] Search engines resemble the online catalog, as both accept search terms and then yield results. Users mistakenly assume that a similar appearance implies a similar trustworthiness and credibility. Yet search

---

[2] This is not to say that library cataloging classification is without its flaws. But it is inconceivable that libraries would ever accept money for promoting an item in their catalog, while search engines do this all the time. On the question of agenda-driven bias it is true that libraries are far from perfect, as cataloging classification schemes arguably embody racist, sexist, and classist biases. Yet the bias found in search engine results is of an entirely different degree, one that libraries would never tolerate. There are reasons for this: first, librarians are paid to make professional judgments, and thus stand at a remove from direct interest. The job of librarians is to make this disinterested judgment, however imperfectly they carry it out. Also, the library profession has a code of ethics that proscribes activities motivated by private interest or profit: "We do not advance private interests at the expense of library users, colleagues, or our employing institutions" (American Library Association, 1995, section VI). Finally, librarians' judgments are for the most part open to scrutiny and criticism by professional peers and are at least potentially open to public debate. By contrast, search engines are founded upon private interest. Although the ethics of corporations may vary, the bottom line is that trustworthy and relevant results are a means to draw users to their advertising, not an end in themselves. Search engines rely in part on 'lay indexers' (to be described below), people whose recommendations will often be motivated by immediate private interest. Finally, because of search engine secrecy, oversight is limited to the corporation, which might act to curb bias only if profit is threatened.

So although libraries are not beyond reproach, they set a far higher standard than do search engines. For purposes of this paper it seems justified to make a sharp distinction between the relatively unbiased and disinterested judgments that librarians make and the profit and agenda-driven biases present in search engines. For clarity of exposition the remainder of the paper will treat this distinction as if it were absolute, though I ask that the reader bear the foregoing qualifications in mind.

engines do not have the cultural institutions standing behind them to assure their integrity. Certainly there is confusion: many of my students now refer to the library online catalog as '*a search engine*.' As Nissenbaum and Introna observe, "*[m]ost users tend to treat search engine results the way they treat library catalog search results*" (Nissenbaum and Introna, 2004: 16). Moxley et al. are explicit about the transfer of credibility: "*Web search engines trade on the associations its users make between the libraries and information professionals they have always relied upon for similar needs in the past*" (Moxley et al., 2004: 64-65).

Even if the credibility mistakenly attributed to search engines is not directly due to previous experience with library catalogs, it should still be understood as borrowing from the credibility of the print format and the social institutions it represents, including libraries. Put another way, if libraries and other institutions of the print tradition had not been so ethical in the past, it is unlikely that search engine users would be so trusting today. The users' belief that search engines should be ethical is thus partly a residuum of this legacy. Search engines that tamper with results (e.g., by accepting money for including advertising in the search results) are exploiting and undermining the faith that one will be presented with information that is not biased by direct personal interest, a trust that has been built by libraries and other institutions of the printed medium over many years (for a similar point see Moxley et al., 2004: 62).

A Major Concern: Search Engine Secrecy

Elsewhere I have argued that Google's success should be ascribed mostly to ethical rather than technological factors, specifically that Google has brought many library values (or valuable library practices) to the Internet environment (Caufield, 2005). For instance, Google brought the important library value of disinterested cataloguing to the search environment, as it refused (and continues to refuse) to include advertisements in the search results, placing them instead in a separate column that is clearly labeled as such (a practice that one of Marable's participants found "*honorable."* Marable, 2003: 31).

Google's ethics seem exemplary in comparison to the egregious practices of other search engines, and this is surely one reason that "*Google has the trust*"

(Harvey, 2004). Yet despite these positive contributions there is cause for concern, particularly with the secrecy of the ranking processes used by Google and all other commercial search engines.

There are reasons for this secrecy, the most obvious being the need to protect proprietary search methods. If Google were to reveal details of its algorithm, presumably competitors would 'borrow' its techniques and Google would lose its competitive advantage. A second reason for this secrecy is described by Terrence A. Brooks, who points out that traditional information institutions (e.g. libraries)

> "… were devised, built and managed by information professionals. Professionals were known, publicly accessible and held to high standards of ethics. Information professionals, such as librarians, were considered to be operating a public trust with a view to the best interests of society" (Brooks, 2004: 7 of 13).

By contrast, search engines rely on a form of collaborative filtering in which large numbers of people contribute to the effort. As Brooks rightly notes, the Google system (or any automated system) gathers information from web authors, who thus act as "lay indexers" of web content (Brooks, 2004).

That the people responsible for this are usually not professionals is problematic enough; what is worse is that they are often working toward private goals (promotion of their products and agendas) rather than the public good. Thus any automatic ranking system is susceptible to manipulation or "spamming," the deliberate introduction of recommendations that serve the interests of the lay indexers (Neuberger, 2005: 4; Spencer, 2004; Nissenbaum and Introna, 2004: 13-14). The more that spammers know about the ranking system, the more they will be able to manipulate it; if Google reveals exactly how it gathers recommendations and calculates rankings, spammers will be able to work the system all the better. Brooks understands this danger, and so he stresses "… the importance of maintaining the ignorance of lay indexers. [….] Google's only possible survival strategy is to keep its parsing and ranking algorithms absolute secrets" (Brooks, 2004).

This point of view is understandable but, in my view, not sustainable. Certainly Brooks describes the necessity of secrecy for commercial search engines: an end to it would perhaps mean an end to the companies. Yet from a larger social perspective this lack of transparency has enormous implications. What does it mean, when a crucial component of the information infrastructure lacks the transparency that is necessary for accountability, review, or criticism? While Brooks maintains that secrecy is necessary for search engines to function, this secrecy seems to be incompatible with the transparency required of information institutions in a democratic society. [3]

Revisiting the Ethics of Link Analysis: Secrecy Prevents Attribution

When the sources that search engines draw upon are not given credit, what harm is done? The question is difficult enough that I will turn to a philosophical statement of the problem. In Plato's *Republic* (359b-360e), two characters present the argument that it is better to be unjust than to be just. They recite a myth about a ring that renders its wearer invisible and so allows great crimes to be committed with impunity. Reasoning from this example they maintain that any logical person, when granted a complete lack of public accountability, will act unjustly and do what is wrong because it serves their private interest.

In a further refinement of their argument (362d-367e) they ask which is more important, actually doing a good deed or simply having the reputation for doing a good deed? If the action is more important than the reputation, then even if the reputation is transferred to one who does not deserve it there should still be some motivation to do good deeds. But this, they say, is nonsense. No one will continue doing good deeds if the credit goes to others. Hence all that is important is the reputation for doing good. The Socratic response to this argument

---

[3] On this issue there are divergent views. Some, usually citing the necessity of secrecy for private companies, agree with Brooks (e.g., Hinman, 2005: 22). Others, usually pointing toward the public's need for accountability, advocate transparency (e.g., Welp and Machill, 2005: 18; Nissenbaum and Introna, 2004: 23). Rieder (2005: 31) tries to split the difference, suggesting that search engines should be obliged to grant access to their indexes and server farms (but retain secrecy of their algorithms), so that public entities can write and run non-private algorithms.

is, in brief, to consider the good of society as a whole, rather than the good of the individual. From this perspective a different view emerges, a society where political structures ensure that deed and reputation will coincide.

The operations of current search engines involve an analogous form of invisibility. Much of the quality of the current generation of search engines is due to link analysis, a process whereby links from one Web page to another are treated as recommendations. These are then aggregated and, along with other factors, used to rank the results. Generally speaking, the more links pointing to a page, the higher its ranking. What is important here is that these recommendations are borrowed without attribution. [4]

In the early days of the Internet, this practice seemed innocuous, perhaps because there simply was not much information there to borrow and the need to create order was so great. Yet in a more developed web environment the practice has serious implications for traditional information institutions. First, search engines harvest relevance and other values from various sources but do not give credit. So in this sense the source of the good deeds (the meaningful recommendations) is not recognized. Second, the reputation for having produced this good work is transferred from the original source to the search engine. For those who succumb to the myth of automated meaning, the search engine now has the reputation and recognition that the original source deserves. So the argument in this paper is

---

[4] Reider (2005: 29, 31) points out that complete attribution of the sources recommending any single result might not be possible, due to the complexity of the process. While this is probably an overstatement the point is still valid, for it is difficult to imagine how this highly-complex attribution could ever be represented to the user on the screen of search results. So some form of invisibility, at least in the presentation of results to users, seems difficult to avoid. But the remaining issue is that secrecy prevents any possible examination of who deserves credit for the borrowed recommendations. It is true that library subject headings are also without attribution, but this is far more acceptable, since librarians have agreed to share such things, and since the profession receives recognition for this work. By the same token, even though library subject headings are also opaque, this is mitigated in that we at least know they come from relatively disinterested sources (please see footnote 2).

somewhat similar to that advanced by Plato's characters: if there is no recognition of good and bad deeds, if all is hidden behind a cloak of secrecy, then what motivation remains for the good deed of creating meaningful and valuable recommendations?

Where Do Trustworthy Recommendations Come From?

It is a common claim that in this new electronic environment librarians are being "disintermediated," their services replaced by other information services, especially automated ones.

> "A generation ago, reference librarians -- flesh and blood creatures -- were the most powerful search engines on the planet. But the rise of robotic search engines in the mid-1990's has removed the human mediators between researchers and information. Librarians are not so sure they approve" (Achenbach, 2004).

While this is the prevalent view, the contention of this paper is that the work of librarians and other information specialists has been incorporated into search engines far more than has been recognized. Their work has not really been removed from the process; rather it has been hidden.

Where do Google's relevant and trustworthy results come from? The short but misleading answer –the myth– is that they come from an algorithm. But mathematical formulae measure and manipulate quantity; they do not evaluate quality. Computers do not decide the relevance and credibility of resources, humans do.

A clue can be found in the sometimes questionable ethical practices of companies that provide 'search engine optimization.' Among other practices used to boost their rankings in Google, spammers set up 'link farms' that create inbound links to their pages. So long as these remain undetected, search engines count these links as legitimate recommendations. Nor is the practice restricted to link farms, as

> "…bloggers have learned that they can 'game' the system by organizing a campaign in which a large number of blogs mention a URL they wish to boost. This simple technique proves to be an effective method of 'index spamming.' Google has tweaked its algorithm to reduce its effectiveness, but hasn't solved the problem" (Wiggins, 2003).

Of course Google gathers its harvest not only from the spammers' link farms and bloggers' agenda-driven pages, but also from all the rest of the Web, plucking the fruit of meaning, packaging it, and delivering it to the user.

But if link farms produce the spam by which Google results (and hence the information environment) are polluted, whence come the healthy harvests, the valuable recommendations that are not motivated by private profit or agendas?   As the extreme case of an all-advertising search engine illustrates, these disinterested and unbiased results are what users desire and, as the Marable study indicates, these are the kind of results they expect.

Evidence that Search Engines Borrow Recommendations from Trustworthy Sources

Because of the manipulation by unreliable 'lay indexers' who try to bias the results, search engines must actively seek out reliable and disinterested sources. In the early days of the World Wide Web most commercial services created their own subject directories. They did this because they were *"…driven by the desire to increase the number of visitors to the site, thereby attracting more advertising dollars"* (Notess, 1997: 72).   But information services eventually moved away from investing effort in subject directories. The following discussion is meant to illuminate the economics (or ecology) of information that is responsible for this trend.

Google has given some explanations of how their process works. A famous early paper by Brin and Page describes an important part of the operation. The PageRank algorithm counts links; in general, the more links pointing to a given page the higher its ranking.   But PageRank also assigns various weights, so that all inbound links do not count equally:

> *"The reason that PageRank is interesting is that there are many cases where simple citation counting does not correspond to our common sense notion of importance. For example, if the web page has a link off the Yahoo homepage, it may be just one link but it is a very important one. This page should be ranked higher than many pages with more links but from obscure places"* (Page et al., 1999; p. 3).

This is a sensible explanation of a clever process, but we should be clear about what it means: in effect, Google's automatic system 'borrows' the recommendations of Yahoo, and also tries to represent the relevance and credibility of these recommendations by raising them in the rankings. In this fashion Google's rankings will in part draw their value from the work of Yahoo's human editors, but no credit will be given.

What Google's founders proposed in 1999 had, by 2001, become a standard search engine practice:

> *"In order to improve relevancy, search engines have begun to compile potential results into directories that further narrow a user's search to a more specific context. In this way, human intervention can weed out the irrelevant results for a given search…. But for poor, understaffed editing teams, this is a little like sifting sand through a tennis racket. The obvious answer is to staff up the Editing Department, but with the pressure on search companies to deliver actual profits, additional headcount is no way to improve margins"* (Belle, 2001: 49-50).

The solution was this: instead of hiring editors, search engines would mine already-existing directories, using them to increase the value of their search results. The human work that had been invested by others could then be harvested and incorporated into the search results.

Though I have not found any discussion of the topic, it is possible that search engines were harvesting the Yahoo directory to increase the relevance of their results.   But another directory, one created by volunteer editors, was frequently discussed as a means of improving relevance.   This resource is the Open Directory Project [ODP]. *"Today, ODP's freely available index is used by major search engines on AOL, Google, HotBot, and Lycos, among many others"* (Belle, 2001: p. 51). Thus the ODP editors' judgments about credibility, trustworthiness, and relevance began to be incorporated into the results of these search engines. Indeed, it seems that any search engine not drawing upon this resource would be neglecting a recognized contribution to its competitive edge.

It is important to understand that a direct link from Yahoo or the ODP is more valuable precisely because it is presumably a disinterested judgment, probably more authoritative and less motivated by private interest than the links offered by most web

sites. It is also important to understand that this value is mined by search engines and then re-presented in a much more accessible form.

> *"The ODP has very useful information," said Sergey Brin, president and cofounder of Google. "But it's tedious to browse. So we put our technology on top [to make it easier to find relevant results without having to scan through lengthy alphabetized lists of links]"* (Sherman, 2000).

Brin's comment outlines the new ecology of information that has come to dominate the Internet. While human editors provide the essential 'useful information,' search engines will absorb that information and deliver it to the user in a much more convenient form. The automated capture of information practiced by search engines means that potentially all information posted in directories and elsewhere on the Web can be absorbed into search engines. Olsen and Hu (2003) have summed up the situation well: *"Google rides on the back of human screening of web sites."*

Are Libraries an Unacknowledged Source of Trustworthy Recommendations?

As these observations make clear, search engines do not create value but rather harvest it. That is not to say that the people who operate search engines do no work at all. They do not send auto-mated tools out to gather these human judgments from all and sundry; Google does not simply aver-age a mass of decisions to arrive at superior intelli-gence.[5]

---

[5] In *The Wisdom of Crowds* James Surowieki has argued that the averaged decisions of many are almost always superior to the judgment of experts, and he attributes Google's success to this phenome-non (Surowieki 2004: 17, 72, 270). Yet Surowieki's hypothesis, when translated to search engines, is a version of the myth of automated meaning, for although it acknowledges the human sources it implies that they are simply averaged. This does not take into account the conflicting biases and agendas on the Internet, which would render all such averages suspect. In such a situation we need the judgment of professionals.

So although Google describes its PageRank system as 'democratic' (Google, 2004b), it is not (Blanke, 2005: 36). This is partly due to the design of the automated system (described above), but there is another reason. Because spammers actively seek to promote their untrustworthy recommendations, search engines will need to actively identify and draw upon trustworthy sources. Google (or any other reasonable search engine company) will begin its search for meaningful recommendations, not by a random canvas, but by targeting institutions known for their quality of judgment. A fairly old web page maintains that *"[Google] has a built-in bias towards EDU and GOV sites that is a refreshing change from the other dot-com spam-laden search engines"* (*Search Engine World*). This undated article was written in approximately 2000, but it seems eminently plausible that Google would espe-cially wish to harvest the unbiased and trustworthy materials of government and educational institu-tions. It is only logical that these recommendations are less likely to be tainted by private interest and so will, on average, have more trustworthy recom-mendations than web pages from the dot-com domain.

A crucial step in creating a reliable and trustworthy search engine is the initial crawl to build the index. The search engine sends out a "spider," a program that crawls the web by following links from one web page to another; the index is built by copying these pages. The choice of starting pages is crucial be-cause it establishes the basis for all subsequent rankings. In 2002 Monika Henzinger, a Google spokeswoman, gave an interview to the *Journal of Internet Cataloging*. Speaking about the process of bringing order to the Web, she made clear that Google's relevance and trustworthiness are founded upon web pages selected for their high credibility:

> *"Most people don't realize that Google finds most of its pages when our robots crawl the Web and jump from page to page via hyper-links. [….] In general, the crawl starts out with a seed set of pages that we know and trust and we think are good. From there we look at what links are going out from that link"* (Monika Henzinger, quoted in De Groat, 2002: 21).

Again, it is only logical that these are not simply web pages, but instead the pages of institutions known to promulgate the most trustworthy and reliable information. The Google representative does not say what those web pages are, though later in her interview she confides *"[t]here is something that*

librarians and other expert information professionals could do for us. I would be very curious to get more feedback from librarians. I feel that there is a lot of knowledge and experience that they have and we are not getting enough of that" (Henzinger, quoted in de Groat, 2002: 27).

It is odd that libraries are never mentioned among the resources drawn upon to improve search engine results. After all, a harvest of all library and library-related web sites would likely be larger and of better quality than could be produced from the frequently-mentioned Open Directory Project. Google's stated intentions are consistent with such an activity. In a December 14, 2004 press release Google cofounder Larry Page is quoted as saying "Even before we started Google, we dreamed of making the incredible breadth of information that librarians so lovingly organize searchable online" (Page, quoted by Letwin, 2004).

It would be fairly easy to set up systems that would 'borrow' or expropriate the knowledge, expertise, and good judgment of trusted institutions. Would not library web pages be a suitable basis for an algorithm that would produce credible and trustworthy (as well as relevant) results, perhaps even the best set of 'seed pages' from which a search engine can 'start out'? The same operation could be performed on other public institutions, for example schools, universities, public radio, and public television. It could also be done with those private companies that are known to provide reliable information, such as newspapers or medical institutions. To the extent that sources of this kind play an important role in Google's search process, Google is borrowing relevance and trust from libraries and other trusted authorities, what Jose-Marie Griffiths has called "the knowledge validating institutions" (Griffiths, 2000: 47).

Is There Surreptitious Gathering of Recommendations from Expert Searchers?

It is well established that search engines monitor user behavior in order to improve performance. Even in the early days of search, "[t]he Yahoo team quickly realized the value of its users' clickstreams. 'People came to our servers and they'd leave tracks,' Koogle says. 'We could see every day exactly what people thought was important on the Internet.'" (Battelle 2005: 62, quoting Tim Koogle, Yahoo CEO). Yahoo used this information to build its

directories, and by 1999 Brin and Page were apparently using the same kind of data to check the accuracy of the PageRank algorithm (Page et al., 1999: 13).

Other search engines have recently announced that they are gathering user information for similar purposes. For instance, Mooter monitors its users and adjusts search results on the fly (Roush, 2004: 39), and SNAP is "…a new breed of search engine that ranks sites by factors such as how many times they have been clicked on, among other things" (Battelle, 2005: 121; see also Walker, 2004). In general, "[s]earch engines keep voluminous logs of user interactions, mainly to divine patterns to make their engines more efficient and more profitable" (Battelle, 2005: 196). Google certainly monitors users: "'the system that has the most users benefits the most,' said Nancy Blachman, a computer scientist and author of an independent guide to using Google (www.googleguide.com). 'Microsoft faces a tremendous challenge because Google fine-tunes its system by watching how users adjust their queries'" (Markoff, 2004).

Most observers seem to think this monitoring is unobjectionable, worrying only about possible privacy violations. That certainly is a concern. But another implication has gone unnoticed, namely that search engines are harvesting the judgments of individual searchers.

We have already seen that Google does not wish to draw its recommendations from websites at random, but rather seeks out the most trustworthy ones. The same logic would apply to the 'recommendations' made by searchers, some of whom will be more trustworthy and otherwise valuable. Google and other search engines are able to identify IP address and more (Hinman, 2005: 23), and the cookie it deposits on each machine can record search histories (Himma, 2004: 71). It would not be difficult for Google (or any other search engine) to identify people who perform a high number of searches, and this could even be correlated with library IP addresses. Certainly there is a threshold high enough to exclude most casual users, yet keep, for instance, the computer at a busy reference desk. The selections made by these expert users could then be boosted in the rankings, even though librarians and others would not be aware that they were contributing to the authority and relevance of Google's results. Nor does there seem to be anything in Google's current privacy policy that would prevent such 'borrowing':

> *"Google collects limited non-personally iden-
> tifying information your browser makes
> available whenever you visit a website. This
> log information includes your Internet Pro-
> tocol address, browser type, browser lan-
> guage, the date and time of your query and
> one or more cookies that may uniquely
> identify your browser. We use this informa-
> tion to operate, develop and improve our
> services"* (Google, 2004a).

Due to secrecy the extent of such borrowing is
unknown. Still, it is an important ethical issue, one
that concerns potentially all searchers, not just
information professionals.   Because the Google
cookie can report search histories for individual
machines over time, intensive research would be an
especially appealing target.   A searcher's choices
could be used to improve the results, which would
then be delivered to the next person with a similar
search, possibly a competitor.   This is clearly an
issue of intellectual propriety, another problematic
aspect of the unannounced 'sharing' that this paper
makes thematic.

Mining and Undermining the Public Resources: The New Ecology of Search

By about 2003 we find a general recognition that
search engines have once and for all outstripped
directories in terms of usefulness:

> *"Yahoo isn't the only directory facing criti-
> cism these days. Search engine marketers
> also point to the Open Directory Project
> (ODP) as an example of how far directories
> have fallen behind algorithmic search pro-
> viders, both in terms of the reach and quali-
> ty of the results they provide"* (Olsen and
> Hu, 2003).

Oddly enough, no one other than Olsen and Hu
seems to have examined why the semantic aggrega-
tors (search engines) are better than the directories.
The assumption seems to be that technology has
replaced human effort, but this is a version of the
myth of automated meaning, as a meaning-
aggregating machine will never replace a meaning-
creating human.

We can outline the ecological configuration of this
new Internet information environment. What does
it mean to say that *"Google rides on the back of*

human screening of web sites" (Olsen and Hu,
2003)? The contents of any directory are public and
are thus available for harvesting by search engines,
which can then 'borrow' the relevant and trustwor-
thy recommendations they find there and include
them in their results. The reverse process is no-
where near as easy, as it requires human operators
to first use the search engine, judge the results, and
then place their choices into a directory or web
page. Perversely enough, any humanly-produced
refinement will once again be 'borrowed' the next
time the search engine crawls the web and updates
its index. And, as already noted, we can add the
corollary that search engines 'ride on the backs' of
individual searchers, as the selections made by
search engine users can be monitored, turned into
recommendations, and then incorporated into the
ranking process.

The euphemistic language used to describe these
activities is 'harvesting' or 'mining,' but these terms
connote productive human activities that extract
value from the non-human world. The systematic
gathering of recommendations by search engines is,
by contrast, an operation that divorces these values
from the individual and institutional sources that
produce them. The library profession would seem
to be especially vulnerable to this unannounced
transfer of knowledge on account of the quality and
quantity of recommendations it makes publicly
available, and because this activity is among the
core services of the profession.[6]

Comparison of Library and Search Engine Cultures

How much the information environment has
changed! It is frequently noted that libraries are not
subject to competitive pressures, and often this is
spoken of as a negative thing. As the sole pur-
veyors of a certain kind of information, libraries
have occupied a quasi-monopolistic position and
thus did not need to pay attention to users' conven-
ience.   The recent ascendancy of search engines
demonstrates the folly of this shortsightedness.

---

[6] Of course other entities provide the public with
recommendations—for instance critics, reviewers,
and bookstores. This paper focuses on libraries
because search engines are not liable to draw
extensively from these other sources, nor are search
engines conceivably in competition with them.

But there are values other than ease of access, and in regard to these libraries are generous. The lack of competition has positive aspects, allowing traditional library culture to be collaborative and tolerant of considerable borrowing between its members. One reason for this is that librarians share a common goal –service to the users, often the public in general– which is more completely achieved by cooperation. The tendency has been to treat this kind of information as a public good. Save for limited indexing and abstracting services, most of this information about information was never subject to market competition, and libraries had no need to guard the values they produced. But while the work of librarians is usually made freely and publicly available, search engines are secretive. The contrast illustrates a fundamental change occurring in our information culture.

Because the search engine is more comprehensive (it harvests many directories and web pages), and because it delivers more convenient results, users naturally flock to this new point of access. Since there is no attribution, users come to believe in the myth of automated meaning, failing to recognize that other sources produce this meaning. So it is somewhat ironic to hear that directories

> "… are being displaced by algorithmic search tools and commercial services that many people… now believe do a better job satisfying Web surfers and advertisers. The transformation is bringing to an end an altruistic era of human editors, who once wielded significant clout in driving traffic to web sites through recommendations made without regard for commercial considerations" (Olsen and Hu, 2003).

In the same way it is somewhat ironic that the founders of Google would cite a human-powered directory as a source of useful information: ironic, because the links are valuable precisely because they have been judged by professional editors. The problem for the Yahoo directory (as with the Open Directory Project and all library web pages) is that it is open to be mined by all search engines, and thus equally useful to Yahoo's competitors. Despite this, Yahoo is in no position to complain, as it practices the same kind of mining.[7]

Whether it is due to users' preference for ranked results, or because directories can be mined by competitors, or simply because public sources can replace much of the work done by paid employees, it seems that Yahoo has begun to limit the work it invests in creating recommendations: *"At least one search engine marketer has said that Yahoo has scaled back on its directory editors slowly over recent months, giving people new duties or emphasizing paid search listings"* (Olsen and Hu, 2003). A related observation is made by Joseph Janes, who asks *"Why Is the Internet Public Library Broke?"* His answer is that *"[t]he IPL is a really great idea, but the bottom line is that it's not anybody's responsibility to fund a 'public library' for the entire Internet"* (Janes, 2003). This is especially true if any quality recommendations published on the Internet will immediately be siphoned off and repackaged into search engine results. The users (and administrators) must think, if search engine results are more copious and convenient, then why do we need a directory?

It is also ironic to hear the concern, common since the advent of search engines, that librarians might be 'losing their relevance.' This usually refers to the fear that librarians will lose their reputation as the best source for meaningful recommendations, so the public (or students) will no longer seek them out. Yet if the thesis of this paper is even partially correct, librarians have been 'losing their relevance' (and their credibility and trustworthiness) in a much more direct and literal fashion, as their web pages and even their individual searches are harvested by search engines.

Discussion

The distinction used throughout this paper is overly-simplistic but nevertheless useful.[8] Google and other

---

[7] The question as to how much can be 'borrowed' from commercial sites needs fuller investigation. It is possible to block search engine spiders from a

website by using a robot.txt file, so commercial sites probably can, for the most part, prevent this form of harvesting. But even if private commercial sites can block spiders and other forms of borrowing, that does not weaken the thesis of this paper, as public sites (including libraries') remain open.

[8] As with all simplifications it is also a falsification and tends toward extremes. Library classification and other services are not so disinterested as herein portrayed (see ftnt. 2). Likewise I simply point to

search engines gather recommendations from two kinds of sources: one has a private interest at stake and is looking to influence search results, making those recommendations biased and untrustworthy; the other is performing a public service by making disinterested judgments about what is most relevant, credible, trustworthy, educational, useful, etc. Search engines remove from these actors all public recognition of the ethical quality of their actions, making them invisible. The reward for the unscrupulous spammer is substantial. Conversely, the reward for those whose disinterested evaluations serve the public good is small to none. This is especially true when the credit for their good deeds has now, under the myth of automated meeting, been transferred to the search engine.

When search engines draw upon this public good and make it appear as a free gift from their private company they are obscuring the source of value. Why would the public support costly institutions such as libraries, when a private entity delivers many of the same things for free, with much greater ease and abundance? Why would a government official or public administrator support public organizations that create order on the Internet? Hiding the sources that create public knowledge encourages disinvestment in them.

Consider an alternative information ecology: if it were not for the semantic aggregators, the Internet Public Library would most likely be flourishing. Does Google do a better job than the directories do? Well, in a way yes, but only because Google draws on all the resources that these directories make public.

Indeed, this seems to be a one-way information economy, where all good is transferred from the creators of meaning to the harvesters of meaning, with no recognition of human labor. This disincentive to invest in the creation of meaningful recommendations, and the corresponding emphasis on harvesting meaning from existing resources, is, I think, pernicious. It is, unfortunately, a major force shaping the information ecology today.

This is not to say that there should not be search engines. Certainly there is value in the work that Google or any competent search engine company

---

the implications of search engines' borrowing from libraries. These simplistic claims are meant to delineate the basic issues at stake, while the reality is more a matter of degree.

does. Librarians know that selecting and organizing resources requires work, and this remains true even when much of the process is automated. Nor is this to say that the semantic aggregation (collaborative filtering) that Google and other search engines perform is inherently bad. On the contrary, it is a very good thing indeed. The point here is simply that some forms of collaborative filtering unfairly distribute the labor and the rewards for that good work. This is particularly true when the library culture of openness and collaboration meets the search engine culture of secrecy and competition. To the extent that the meaning produced by libraries is siphoned off and repackaged by search engines (which only the search engines know), to that extent libraries are collaborating with their competitors.[9]

Conclusion

Many reasons have been given for holding search engines to high ethical standards. Most of these have to do with the essential services search engines provide to the public. This paper adds that search engines are ethically obligated because they draw so much from public sources. To some large extent the intelligence they display is not their own, but is rather borrowed. In a less direct fashion, the trust that users extend to search engines is also borrowed, for it has been built by the ethical practices of other information institutions. For these reasons search engines' ethical obligations are far greater than previously recognized.

I have argued that libraries are especially vulnerable to this borrowing, not because I know that search engines draw extensively upon their work (secrecy prevents that), but because it is logical that they would.

It seems we have entered a new ecology of information, one where search engines capture and recycle all meaningful recommendations. After this content has been repackaged by search engines, the

---

[9] Obviously much could be said regarding the Google Print project and other collaborative efforts. Suffice it for now to quote Rory Letwin, who notes that it "… is not a victory for libraries if the real meaning of this development is simply the transfer of all this information out of our humanistic institution and into the marketplace" (Letwin, 2004).

original sources of meaning appear irrelevant and unnecessary. So even in the unlikely scenario where Google or another search engine creates an ideal aggregation system, one that gives appropriate weight to the disinterested judgment of public institutions and avoids all the corruptions that lay indexers can introduce, even this ideal system will still contribute to the corruption of the information infrastructure. It will do so by mining, exploiting, and ultimately undermining support for all publicly-motivated efforts. Because search engines provide no acknowledgement of these sources they encourage the myth of automated meaning, the belief that these things are produced by technology alone and could be provided without the efforts of people and institutions. If this myth comes to be universally believed, the institutions that create these valuable recommendations will decline. Then much of the trustworthy information that users expect will no longer be there for search engines to harvest, and the world will be an intellectually poorer place.

References

Achenbach, Joel: "Search for Tomorrow," Washington Post, Sunday, February 15, 2004, p. D01.

American Library Association, "Code of Ethics" http://www.ala.org/ala/oif/statementspols/codeofethics/codeethics.htm. (1995) Last accessed March 6, 2006.

Battelle, John: The Search: How Google and Its Rivals Rewrote the Rules of Business and Transformed Our Culture. New York, Portfolio (Penguin Group) 2005.

Belle, Jeff: "ODP Shakes Up Search Directory Content Revenues". EContent. February/March 2001, 49-52.

Blanke, Tobias: "Ethical Subjectification and Search Engines: Ethics Reconsidered," International Review of Information Ethics, vol. 3 (June 2005), pp. 34-38.

Brooks, T. A.: "The Nature of Meaning in the Age of Google" Information Research, 9 (3) paper 180 (2004) http://InformationR.net/ir/9-3/paper180.html

Caufield, James: "Where Did Google Get Its Value?" portal: Libraries and the Academy, vol. 5, no. 4 (October 2005), pp. 555-572.

De Groat, Greta: "Perspectives on the Web and Google: Monika Henzinger, Director of Research, Google," Journal of Internet Cataloging, vol. 5, no. 1, 2002, pp. 17-28.

Fallows, Deborah: "Search Engine Users: Internet Searchers Are Confident, Satisfied and Trusting -- but They Are Also Unaware and Naive," (Pew Internet & American Life Project), January 23, 2005. Available at http://www.pewinternet.org/PPF/r/146/report_display.asp. Last accessed January 12, 2006.

Goh, Dion H., and Rebecca P. Ang: "Relevancy Rankings: Pay for Performance Search Engines in the Hot Seat," Online Information Review, vol. 27, no. 2, 2003, pp. 87-93.

Google (2004a): "Google Privacy Policy" http://www.google.com/privacy.html. Last accessed September 5, 2005.

Google (2004b): "Our Search: Google Technology," http://www.google.com/technology/. Last accessed September 5, 2005.

Griffiths, Jose-Marie: "Deconstructing Earth's Largest Library," Library Journal, August 2000, pp. 44-47.

Hansell, Saul: "Google and Yahoo Settle Dispute over Search Patent", New York Times, August 10, 2004, p. C6.

Harvey, Fiona: "New Media," Financial Times Limited (London) March 30, 2004, p. 8.

Himma, Kenneth Einar: "Legal, Social, and Ethical Issues," The Internet Encyclopedia, Hossein Bidgoli, Editor-in Chief. Hoboken, New Jersey, John Wiley & Sons Inc. 2004, pp. 464-476.

Hinman, Lawrence M.: "Esse est Indicato in Google: Ethical and Political Issues in Search Engines," International Review of Information Ethics, vol. 3 (June 2005), 20-25.

Janes, Joseph: "Internet Librarian: Why Is the Internet Public Library Broke?" American Libraries, March 2003, p. 86.

Larsen, Elizabeth: "Finders Keepers: How Search Engines Control Our Access To Information," Utne Reader, May/June 1999, pp. 22-3.

Letwin, Rory: "On Google's Monetization of Libraries," Library Juice, Vol. 7 No. 26 (December 17, 2004). Available online at http://libr.org/Juice/issues/vol7/LJ_7.26.html. Last accessed May 25, 2005.

McRae, Hamish: "How Google Helped to Change the World," The Independent (London), August 18, 2004, p. 27.

Marable, Leslie: (Consumer Reports WebWatch), "False Oracles: Consumer Reaction to Learning the Truth about How Search Engines Work. Results of an Ethnographic Study." Consumer WebWatch, June 30, 2003. Available online at http://www.consumerwebwatch.org/dynamic/se

arch-report-false-oracles-abstract.cfm. Last accessed May 15, 2005.

Markoff, John: "The Coming Search Wars," The New York Times, February 1, 2004.

Moxley, David, Joni Blake, and Susan Maze: "Web Search Engine Advertising Practices and Their Effect on Library Service," The Bottom Line: Managing Library Finances, vol. 17, no. 2 (2004), pp. 61-65.

Neuberger, Christoph: "Funktionene, Probleme und Regulierung von Suchmachschinen im Internet," International Review of Information Ethics, vol. 3 (June 2005), pp. 3-17.

Nissenbaum, Helen and Lucas D. Introna: "Shaping the Web: Why the Politics of Search Engines Matter," in The Internet in Public Life (ed. Verna V. Gehring). Oxford, Rowman and Littlefield Publishers, Inc, 2004, pp. 7-28.

Notess, Greg R: "New Databases from the Internet Search Services," Database Magazine, vol. 20, no. 2 (April/May 1997), pp. 72-75.

Olsen, Stephanie, and Jim Hu: "The Changing Face of Search Engines", CNET News.com, March 24, 2003. Available at http://news.com.com/2100-1032-993677.html?tag=fd_lede1_hed. Last accessed May 9, 2005.

Page, Larry, et. al.: "The PageRank Citation Ranking: Bringing Order to the Web", November 11, 1999 http://dbpubs.stanford.edu:8090/pub/1999-66. Last accessed October 2, 2005.

Plato: The Republic, (Paul Shorey, translator), Cambridge, Massachusetts, Harvard University Press 1982.

Rieder Bernhard: "Networked Control: Search Engines and the Symmetry of Confidence," International Review of Information Ethics, vol. 3 (June 2005), pp. 26-32.

Roush, Wade: "Search Beyond Google," Technology Review, vol. 107, no. 2 (March 2004), http://www.technologyreview.com/articles/04/03/roush0304.2asp. Last accessed July 17, 2005.

Scott, David M.: "Paying Your Way to the Top," Econtent, May 2003, pp. 32-38.

Search Engine World: "Search Engine History Part 2" (undated). http://www.searchengineworld.com/engine/players2.htm. Last accessed January 6, 2005.

Sherman, Chris: "Google Introduces Web Directory Using Netscape's Open Directory Project Data." Information Today, March 27, 2000. Available at

http://www.infotoday.com/newsbreaks/nb000327-1.htm . Last accessed May 14, 2005.

Spencer, Stephan: "Casing the Competition," Catalog Age, April 1, 2004, pp. 35-7.

Surowieki, James: The Wisdom of Crowds: Why the Many Are Smarter than the Few and How Collective Wisdom Shapes Business, Economies, Societies, and Nations. New York, Doubleday 2004.

Welp, Carsten, and Marcel Machill: "Code of Conduct: Transparency and the Net: Search Engines," International Review of Information Ethics, vol. 3 (June 2005) p. 18.

Walker, Leslie: "Online Search Universe Is Expanding," The Washington Post, October 21 2004, p. E01.

Wiggins, Richard: "The Privilege of Ranking: Google Plays Ball," Searcher, vol. 11, no. 7 (July/August 2003), p. 23.