Bias, privacy, and personalization on the web

Ian H. Witten
Department of Computer Science, University of Waikato, New Zealand
July 2007

There's been a lot of buzz about how the web (and more generally the Internet) is a public space, and a public good. This is exactly what stimulated rabid investment in web-oriented enterprises during the dot-com boom. The perception is grounded in the hard work, vision, and unwavering commitment of those stalwart pioneers who struggled to keep first the Internet and then the web open, free, and universal. But in accepting the web as a public space we must acknowledge the risks inherent in the ways we access it.

We live in interesting times. The past five or ten years have transformed a situation where most of the information we use has been obtained through referrals—from people we know, links in web pages we browse, or references we consult—into one where we locate most of our information using Internet search engines. The term "web dragons" is an apt metaphor for these portals through which we access society's treasure trove of information. Dragons connote unprecedented power whose source is mysterious and totally unfathomable, combined with some degree of moral ambiguity. In the Orient dragons are wise and wonderful; in European mythology they are dire and dreadful.

The transformation from hyperlink-based surfing to full-text searching has some rather disturbing aspects. One is the need for total, blind reliance on black-box mechanisms whose inner workings are a complete mystery—not just in practice (after all, I don't know much about how my car works) but in principle (I can reverse engineer my car but am prohibited from trying to find out how my search engine works[1]). A second is that web dragons centralize the control of information, which is potentially risky—indeed, potentially explosive. The problems cannot really be addressed by legislation, because search engines do their work for free: how can you complain about a service that gives its product away? And putting a centralized information utility into public rather than private hands is not really likely to help. A third is the dynamics of searching versus surfing: minority pages, admittedly only rarely encountered while surfing, are *never* encountered through searching. When did you last click through to the 1,000,000th search result—or even the 100th? Along with these disturbing trends are many liberating forces: we all use search engines every day, and are immensely grateful to them. Eternally grateful?—perhaps its too soon to say.

Web users are utterly dependent upon their search tool. They can choose their query terms but have no control at all over the strategy the search engine adopts. For instance, all pages change their rank at unpredictable times when search engines update their index and algorithms. And while the dragons could (at least in principle) analyse the consequences of their actions, the rest of us have no way of doing so

[1] For example, Google's terms of reference state that "You may not (and you may not permit anyone else to) copy, modify, create a derivative work of, reverse engineer, decompile or otherwise attempt to extract the source code of the Software or any part thereof."

because the basis of their decisions is secret. For one thing, it's closely guarded commercially confidential information—but the problem runs far deeper than that. If the dragons' algorithms were known they could be exploited by people to manipulate the search results and bring certain pages to the top. The economic value in having your pages appear first in response to relevant searches is immense. The dragons' inner workings must be kept under wraps in order to combat web spam. This gives them the power to transform the perceived reality of the web arbitrarily, unilaterally, and without any notice or comment. Even if you have some inkling what is going on behind the scenes today you have no way of predicting what might happen tomorrow.

Users place blind trust in their search results, as though they represented some kind of objective reality. They hardly notice the occasional seismic shifts in the world beneath their feet. They feel solidly in touch with their information, blissfully unaware of the instability of the mechanisms that underlie search. For them the dragons are omniscient. And, by the way, it's not just the web. Search engines are taking over our literature. Depending on how the copyright issues—which are a bone of much contention—play out, the very same dragons may end up controlling all our information, including the treasury of literature held in libraries. The problems of bias, privacy, and personalization that are identified in this article transcend the World-Wide Web as we know it today.

How search engines work

The first generation of search engines worked by counting words, weighing them, and measuring how well each document matches the user's query. This was an appropriate, familiar, and scientific way of dealing with the objective reality represented by a set of documents, and one that we can all understand.

Today, search engines count links as well as words and weigh them too. For each page a number (often called *PageRank*) between 0 and 1 is calculated that indicates its weight, or prestige. Pages gain prestige from every page that contains a hyperlink to them, and bestow it on every page to which they link. More accurately, a page's prestige is *shared* between all pages that it points to—this ensures that prestige cannot be artificially manufactured simply by populating a page with a plethora of outgoing hyperlinks.

The prestige of a page is the sum of the prestige of all pages that point to it, each one being divided by the number of out-links from that page. This involves a certain circularity, and without further analysis it's not clear that it can be made to work. But it can. The calculation can be formulated as a vast system of linear equations, one for every web page, which is solved to give a number between 0 and 1 to each page as its "prestige." Artefacts like broken links and circular references create anomalies in this system of equations, but these can be dealt with fairly easily.

Contrast the methods used by early search engines and today's. Both are objective: the first measures properties of individual documents; the second measures properties of the web as a whole. However, today's dragons do not divulge the recipe they use to weigh and combine links and words. We cannot know it, and are not allowed to: it's a trade secret. More fundamentally, secrecy is an unavoidable side effect of the need to maintain the illusion of an objective reality that the bad guys—spammers—are trying to distort. Even were it not commercially confidential, if search engines were "open source," their precise mechanism would still need to be concealed to defend against distortion of the results by the bad guys.

How we use them

Experienced searchers exercise great discrimination in how they search the web—or at least they know they ought to. They often consult more than one search system, including the many specialized tools that are available. They readily distinguish advertising from third-party opinion, and they evaluate and cross-check the source of opinions. They always carefully assess the credibility of the pages that are returned, using knowledge and experience built up over time. But most users—particularly inexperienced ones—access the web using just one search portal and accept what it returns on good faith. If they are dissatisfied with the result of their query the overwhelming majority prefer to formulate another query for the same engine than switch to another information portal. Ordinary users do not realize that they lack any knowledge of how information is being selected for their attention—or if they do, they rarely reflect upon this fact.

Surveys have revealed that over two-thirds of users believe that search engines are a fair and unbiased source of information. In spite of the trust they place in these tools, the most confident users are ones that are less knowledgeable and experienced in the world of search. In particular, many are blissfully unaware of two controversial features: commercialism, in the form of sponsored links, and privacy, because search engines track each user's search history—and under certain circumstances, their browsing history too.

Studies have shown that only around 60% of users can identify commercially sponsored links in the search results, a proportion that has remained unchanged over the past two years. Ignorance of potential privacy invasion is even more prevalent. Nearly 60% of users are unaware that their online searches are tracked, and, when informed, over half disapprove of this practice. Some claim they would even stop using a search engine if they knew.

The potential effect of commercial—and political—exploitation of individuals' search history is dramatic. For the sake of democracy and transparency in our society, people's attention must be drawn to the possibility that their privacy may be violated. Most users remain unaware of the processes they invoke when interrogating the web. As citizens and consumers, we all have the right to know what is happening, who is in a position to exploit our private data and what are the guarantees that the services we use are fair and unbiased.

An example

The web contains many inbuilt biases. As a concrete example, consider the information about different countries that is obtained by simply submitting their name to a standard search portal. These are certainly not well-focused queries, but you can imagine citizens casually seeking general information about their homeland, or enquiries from potential tourists. We choose this modest example not for its subtlety because it is something to which we can all relate.

Table 1 shows the top five links returned by a search engine (Google) in early 2006 for the queries *United States*, *United Kingdom*, *South Africa*, and *New Zealand*. Of course, search results are highly volatile; they will certainly have changed radically by the time you read this—as we will see below. Nevertheless, they make a clear point. The results for the first two countries largely reflect their citizens' interests: four of the five links are to national institutions. For the last two they largely

reflect visitor and immigration information: only one link each is to a national institution of central interest to citizens. Moreover, the *CIA World Factbook* figures prominently in three of the four results, a fact that these country's citizens may not appreciate—they could be forgiven for assuming that it presents a U.S.-centric view.

United States	***United Kingdom***	***South Africa***	***New Zealand***
States and Capitals	CIA World Factbook—United Kingdom	News results for "South Africa"	The official tourism New Zealand site
US Senate	UK—National Statistics	Welcome to South Africa	New Zealand Herald
US Census Bureau	Patent Office of the UK	CIA World Factbook—South Africa	CIA World Factbook—New Zealand
US Government Official Web Portal	UK Parliament	South African Government Portal	National Library of New Zealand
US Postal Service	Website of the UK Government	South Africa Online (tourism)	Immigration New Zealand

Table 1 Top search results for four different countries (Google, early 2006)

Table 2 shows the top five links returned (by Google) in July 2007; this time we have included *India* alongside the other four countries. Thankfully the CIA World Factbook has been demoted (to position 26, 16, 25, 20 for *United Kingdom*, *South Africa*, *New Zealand* and *India* respectively), but it is replaced by Wikipedia—perhaps an equally controversial information source. There has been some movement towards a more equitable distribution of information returned for different countries. For example, Wikipedia is also the top hit for *United States*. For all five countries, official government portals now appear in the top five hits. One tourism site—an official Government one—now appears for the UK. However, commercial sites figure strongly for *South Africa*, *New Zealand* and *India*. Tourism still dominates *New Zealand* (3 out of 5 hits), features strongly for *South Africa* and *India* (2 out of 5 hits), and is entirely absent for *United States*.

United States	***United Kingdom***	***South Africa***	***New Zealand***	***India***
United States – Wikipedia	United Kingdom - Wikipedia	Welcome to South Africa (national tourism site)	The official tourism New Zealand site	India - Wikipedia
States and Capitals	VisitBritain (national tourism agency)	South Africa - Wikipedia	New Zealand - Wikipedia	Welcome to India (commercial tourism website)
US Government Web Portal	UK Government department for foreign affairs	South Africa hotels … (commercial tourism site)	100% Pure New Zealand (commercial tourism site)	Incredible India (Govt tourism website)
US News	UK history, geography, government, and culture	South Africa's official gateway	Immigration New Zealand	Yahoo! India
US history, geography, government, and culture	UK Indymedia: independent media organizations	South Africa Government Portal	New Zealand travel (Govt tourism website)	National Portal of India

Table 2 Top search results for five different countries (Google, mid 2007)

Bias

It is hard for us to appreciate the inbuilt biases caused by unequal access to the web. Note that these biases are subtle and our example is not; we use nations merely as a simple, easily-graspable illustration. We are certainly not suggesting nationalistic solutions. Indeed, we would argue strongly against them. Enterprises organized on a national or regional scale with a component of public leadership and funding are a far cry from the lone young geniuses, working for love rather than money, who created the search engines we have today and grew into talented entrepreneurs whose dragons are breathing fire at the advertising legends of Madison Avenue. The efforts of national governments are most unlikely to lead to better search. Anyway, the problem is a far broader one of multiple perspectives in general.

The issue is both complex and slippery. Search engines act according to legitimate commercial interests when they privilege certain mainstream results. In doing so they also satisfy the desires of most users, who are primarily interested in information from major web sites. But a direct consequence of the legitimate behaviour of private actors is a shrinkage in the public space. In the long run everyone loses—including search engines, whose popularity is founded on a collectively shared belief that they provide fair and equitable access to the full extent of the riches contained in the largest information repository on earth.

When we search the web we seek more than an answer to a question: we also strive to determine what we do not know. As Socrates asked 2,400 years ago, how can you tell when you have arrived at the truth when you don't know what the truth is? John Battelle, an influential commentator who founded the trendy technology magazine *Wired* and has personally interviewed many prominent figures in the search business, recently identified two reasons for searching online: to recover things that we know exist on the web, and to discover things we assume must be there. In the first case, when trying to recover something we know exists, we will likely recognize the effectiveness (or lack of it) of the response to our query—for the process is one of recollection, not discovery. In the second, he has rediscovered Socrates' paradox: it will be far from easy to evaluate the results received. We can welcome the information that the search engine provides, or reject it; but either way we can do no more than guess. Most likely we will accept the result, for with no clue about what to expect how can we reject the proposed information on the basis of quality?

In practice, many users exhibit an acute lack of awareness when evaluating sources thrown up by their web queries. A study of college students who used search engines to answer a set of questions found that they uncritically accepted their responses. Subjects placed full reliance on information presented by the web, and had complete confidence in search engines as the privileged way to access it. In the fields of advertising, government affairs and propaganda students were particularly susceptible to misinformation and came up with incorrect answers. Clearly, users require training in ways of evaluating information sources, and in the need to reflect critically on the results yielded by any given query. Search engines should be no more than a starting point for the complex process of research and evaluation. For the web to remain a public good, the public—not just students, but the populace at large—must be trained to use it discriminatingly.

Privacy

Most major web sites publish privacy policies, but often only in small print that is hard to find. If you do have the patience to locate and read them you will discover that popular sites have policies that allow them to do anything they want with the personal data you give them. This means that the owners are prepared to share personal information with third parties whenever—in their own opinion—they need to, without having to inform users at all. You would never know; you would never know why; and you would have no appeal.

On the other hand, when asked to register on a website users freely donate their personal information without reflecting on whether or why the requested information is required. There's little point in worrying about such matters because you have no opportunity to negotiate or question what is being asked for: the choice is simply to proceed with the registration process, or not. Of course, there is no compulsion to use any web site: users benefit from an information service for which no charge is made.

The services provided by web dragons are hardly optional in today's world of information. Without search engines knowledge workers would be crippled. And although you may not have to explicitly register for a search service, web query data is a marketer's dream. (It's also a blackmailer's dream, a private investigator's dream, and a nosy government's dream.) This points the spotlight at the web dragons' privacy policies, and raises questions about exactly what is meant or implied by every word and clause.

Ethical considerations of online privacy are governed by two separate principles. The first, *user predictability*, delimits the reasonable expectations of a person about how his or her personal data will be processed, and the objectives to which the processing will be applied. It is widely accepted that before people make a decision to provide personal information they have a right to know how it will be used and what it will be used for, what steps will be taken to protect its confidentiality and integrity, what the consequences of supplying or withholding the information are, and any rights of redress they may have. The second principle, *social justifiability*, holds that some data processing is justifiable as a social activity even when subjects have not expressly consented to it. This does not include the processing of sensitive data, which always needs the owner's explicit consent.

In the context of web search, it is frequently the case that an individual's query stream can be used to identify whom that person is. The dragons know who we are—or can easily find out. Do their privacy statements respect the principles of user predictability and social justifiability? Hardly. Perhaps the problem stems from the cost-free nature of the service, and in future users who are concerned about privacy might be able to have it—at a price.

In addition to searching the public web, there are tools for searching your private file space. The dragons offer downloadable desktop utilities with which you can search your files and the web at the same time, using exactly the same interface. This exploits an amazing weakness in computer operating systems: until recently it has been far easier to find information on the web at large than in your own files! Of course, conjoint searching further threatens the distinction between public and private information, for in order to offer such services the dragons' programs obviously have to access your private files.

There are many other threats to online privacy. Social software stores, aggregates, and organizes user information and preferences. Some sites encourage people to store and share their web bookmarks. Others let surfers store the web pages they are interested in, revealing to the program their entire clickstreams and their selection of online documents. Still others store your digital photographs and videos for free, with no space restrictions, providing you agree that others can see them. These systems offer useful and amusing services, but require users to renounce privacy in favour of either the service provider or the world at large. The world at large, of course, includes the service provider, who has privileged opportunities for data aggregation.

Users will collectively determine whether personalized web systems and other social software turn out to be a success. Regardless of the outcome, it is clear that private spaces are progressively being eroded. Traditional views on privacy are being supplanted by a new world in which people trade personal information for free access to tools that help manage the complexity of online life. You can choose to forsake either your privacy or the convenience of these tools. This raises questions that do not have ready-made solutions.

Anonymity, privacy and security are amongst the most important social issues raised by today's ubiquitous use of the web—and the most difficult to provide any guarantees stronger than the "good faith" claims of the major portals. If you do not trust the dragons, you should not use them. And you need to trust not just them but their political masters, the governments and regimes in which they operate. Not only today but all the way into the distant future, when your every act may be exhumed and subjected to hostile scrutiny. In our uncertain world, rife with social and political unease, how can anyone do that?

Personalization

Many of us assume that the only thing needing protection is intimate and sensitive information within the private sphere. We might even go so far as to claim that there is a realm of public information about persons to which no privacy norms apply, or that aggregating information does not violate privacy if the parts, taken individually, do not. But both are wrong. Just because an event occurs in public does not imply that it automatically belongs to the public sphere. The fact that a rape took place in Central Park does not justify the victim being interviewed by the media in order to inform the public about what happened. In a messy divorce a couple's private affairs are paraded in front of the judge in a public courtroom open to everyone, but this openness is not sufficient reason to publish the transcript on the Internet where it can be located by querying search engines.

As for the second assumption, when pieces of information are aggregated, compiled and assembled, they can collectively invade privacy even though taken individually they do not. You can use a search engine to find out about your next date, the candidates for tomorrow's job interview, your boss's résumé. Whatever we discover we are then prepared to consider as that person's identity. Though powerful and informative, this is so intrusive as to constitute a serious invasion of privacy—even though everything online is public. The act of aggregation introduces bias, and could add further information or misinformation. Suppose you produce a personal profile on someone from information on the web. You will almost certainly, for purely pragmatic reasons, be strongly influenced by the order of search results for the subject's name. Yet while not entirely arbitrary, this order is probably mostly

irrelevant for finding suitable information to include in the profile. The profile is biased—quite apart from any inaccuracies in the information being compiled.

Efficient and effective methods of communicating information are a wonderful thing. But they have a flip side. People have a right to privacy, a right to control the balance between their public and private personae. Whereas you can make purchases anonymously by paying cash and refusing to participate in the supermarket's loyalty card scheme, you cannot conceal your identity so easily when shopping online—and therefore leave yourself open to junk e-mail. If you teach a university course, related information may appear on the institutional website—including your e-mail address. You may wish to share this private information with students, but not give it to the world. But to exploit the possibilities offered by the network to communicate with your students, you have to accept the risk of your address appearing in spammers' databases.

The pervasive intrusion of the Internet into all aspects of our lives muddies the distinction between an individual's private and public space. Some liken the web to a kind of universal library that contains all recorded knowledge. But here's the difference: the web is not just a (potential) record of all external knowledge, but a (potential) record of all personal information (and misinformation) too, information about our e-mails and interests, our every word and action. Personal information, or what purports to be personal information, can be merged and assembled in meaningful and meaningless ways. The web dragons are not just the high-priest librarians who mediate our access to the world of knowledge. They are the friends, counsellors, and tribunals that mediate our access to society too.

Towards solutions

So far we have examined critical issues that affect the web and how we use it: issues of bias, privacy, and personalization. Now it is time to reflect upon possible solutions to the problems we have raised.

Bias. Bias can only be addressed by recognizing the importance of communities and giving them an explicit role in determining the prestige of web pages, and hence the ordering of search results. We all belong to communities. In real life we want our communities to be open and transparent: we want to understand and participate in the processes of membership and governance. We recognize that one size certainly does *not* fit all. And one of the great things about the web is that it's full of communities. The group affairs are the fastest-growing parts, and there's a plethora of different ways of organizing them. Some are anonymous, some pseudonymous. Some are moderated, others immoderate. Some require special qualifications to join; others are open. Some recognize tribal elders; others favour equality. Some have multiple tiers of members: serfs, commoners, lords and ladies, royalty—or in contemporary terminology, lurkers, contributors, moderators, gurus.

Yet today's search engines are blind to all this. Eyes averted, they treat the web as objective reality, not as a social organism. They fail to recognize their users as social creatures who want to work and play within communities—not within some gargantuan hollow-echoing info-warehouse. In order to fix problems of spam, they make decisions that discriminate against certain pages, certain web sites. They make these decisions in the interests of users, on behalf of the community. Most likely they are very good decisions—none of us condones child pornography, or blatant commercialism, or misuse of resources. But I believe that this is not their job, that

they should keep out of the socio-political business of determining, and imposing, community norms. Such decisions should arise out of the community and not be dictated from above.

The way the dragons deal with spam is by imposing a single worldview on the web. But spam is just the tip of the iceberg. In truth there are many, many communities, each entitled to its own point of view, its own values, its own set of prestige values for each page that will determine how prominently they will figure in the search results. The dragons should not be involved in defining communities, or facilitating them, or meddling with them. They should simply recognize them and allow one to search within them. One way of doing this, which most dragons already accommodate, is to restrict search to a particular area of the web, or set of pages. That's simple—and too simplistic. Instead, it would better reflect user needs to restrict the *point of view* to a particular community by computing the prestige of each and every page with respect to a particular set of pages that are specified by the community.

Doing this would allow just the right degree of community participation in the search process. Realistically speaking, users do not really want to know every intricate technical detail of how search is actually made to work. But society should take out of the hands of the dragons decisions about what is appropriate and inappropriate information on which to base judgements about prestige—for example, what is spam and what is not. Future search engines can encourage community involvement without dictating how communities are formed and run. Today's search engines are a first step, an amazing first step, but nevertheless just the beginning.

Privacy. New structures of peer-to-peer networks offer a refreshing alternative to the trend towards centralization that the web dragons exemplify. There are already schemes that pay particular attention to protecting the privacy, security and anonymity of their members. Documents can be produced online and stored in anonymous repositories. Storage can be replicated in ways that guard data from mishap far better than any institutional computer backup policy, no matter how sophisticated. Documents can be split into pieces that are encrypted and stored redundantly in different places to make them highly resistant to any kind of attack, be it physical sabotage of backup tapes, security leaks of sensitive information, or attempts to trace ownership of documents. Your whole country could go down and your files would still be intact.

Peer-to-peer architectures encourage the development of tools that are capable of protecting privacy, resisting censorship, and controlling access. The underlying reason is that distributing the management of information, shunning any kind of central control, really does distribute responsibility—including the responsibility for ensuring integrity and anonymity. There is no single point of failure, no single weakness. Of course, no system is perfect, but the inventors and developers of peer-to-peer architectures are addressing these issues from the very outset, striving to build robust and scalable solutions into the fabric of the network rather than retrofitting them afterwards.

Leading-edge systems guarantee anonymity and also provide a kind of reputation control, which is necessary to restore personal responsibility in an anonymous world. It is hard to imagine how distributing your sensitive information among computers belonging to people you have never met and certainly do not trust can possibly guarantee privacy!—particularly from a coordinated attack. Surely the machines on

the network must whisper secrets to each other, and no matter how quietly they whisper, corrupt system operators can monitor the conversation? The last part is true but the first is not. Strange as it may seem, new techniques of information security guarantee privacy using mathematical techniques. They provide assurances that have a sound theoretical foundation rather than resting on human devices like keeping passwords secret. Even a coordinated attack by a corrupt government with infinite resources at its disposal that has infiltrated every computer on the network, tortured every programmer, and looked inside every single transistor cannot force machines to reveal what is locked up in a mathematical secret. In the weird world of modern encryption, cracking security codes is tantamount to solving puzzles that have stumped the world's best minds for centuries.

In the future standards will be established that allow different peer-to-peer sub-networks to coexist. They will collect content from users and distribute it around in such a way that it remains invisible—mathematically invisible—to other users. In these collective repositories we will, if we wish, be able to share resources with our chosen friends and neighbours, ones who we consider reliable and who have common interests.

And search will change. Search engines may be able to crawl the network but they won't be able to unlock the words in our documents—they won't even be able to patch the fragments of the document together. Of course, much information will be public, unencrypted, and searchable. However, in a world where content is divorced from network structure new strategies will be needed. In keeping with the distributed nature of the information, and in order to preserve scalability, computation will also be distributed.

Personalization. The final issue, personalization, is perhaps the toughest of all. Though extremely useful for users, it is extremely intrusive and, without careful handling of sensitive data, has grave consequences for individual privacy. Personalization is a mixed blessing. On one hand it offers users an interface that has been specially designed to accommodate their preferences and present information customized to their tastes. On the other, users expose themselves to the risk of privacy invasion by making their profile available to others. The risks are determined by the location of the profile—where it is stored and who has access to it. Decentralized peer-to-peer networks will reduce the risks but not eliminate them entirely. The dragons accumulate a vast collection of semi-private, semi-public information, a new treasure of ineffable value acquired with the implicit but unconscious consent of web surfers. Do the advantages that personalization bestows justify this gift? This will be one of the most challenging questions that arise in years to come.

Further reading

The arguments in this article are developed at great length and in more detail in the book *Web Dragons* by Ian H. Witten, Marco Gori and Teresa Numerico, published in 2007 by Morgan Kaufmann.