Journal of Universal Computer Science, vol. 14, no. 10 (2008), 1739-1762
submitted: 28/4/08, accepted: 26/5/08, appeared: 28/5/08 © J.UCS

Searching ... in a Web[1]

Ian H. Witten
(Department of Computer Science, University of Waikato, New Zealand
ihw@cs.waikato.ac.nz)

Abstract: Search engines—"web dragons"—are the portals through which we access society's treasure trove of information. They do not publish the algorithms they use to sort and filter information, yet what they do and how they do it are amongst the most important questions of our time. They deal not just with information *per se*, but evaluate it in order to prioritize it for the user. To do this they assess the prestige of each web page in terms of who links to it. This article explains in non-technical terms what is known about how web search engines work. We describe the dominant way of measuring prestige, relating it to the experience of a surfer condemned to click randomly around the web forever—and also to standard techniques of bibliometric evaluation. We review alternatives: some strive to identify subcommunities of the web; others learn based on implicit user feedback. We also takes a critical look at how people use search engines, and identify issues of bias, privacy, and personalization that crucially affect our world of information today.

Keywords: search engines, web search, PageRank, search bias, privacy, personalization, information ethics
Categories: K.4.2, K.4.0, K.4.1, H.3.3

1 Introduction

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 [Witten 07]. 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 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 also in principle (I can reverse engineer my car but am prohibited from trying to find out how my search engine works[2]). A second is that web dragons centralize the control of information, which is potentially

[1] A version of this paper was presented at SACS 2007 in Graz, Austria on November 6, 2007, see http://www.cs.tugraz.at/sacs

[2] 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."

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 as search engines update their index and algorithms. And while the dragons could (at least in principle) analyze the consequences of their actions, the rest of us have no way of doing so 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, for 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 dragons the power to transform the perceived reality of the web unilaterally, and without any notice or comment. And, from what we know, they do. 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 below transcend the World-Wide Web as we know it today.

This article explains in non-technical terms the techniques on which today's web dragons are based. It does not cover the underlying classical methods of information retrieval, but takes them as given and shows how they have been extended into mechanisms for searching the web. This involves prioritizing information for the user by automatically assessing its prestige, and the dominant technique for doing so is described in Section 3. The next section shows a useful way of looking at prestige in terms of the behavior of a random surfer, which leads to an examination of the large-scale structure of the World-Wide Web. Section 5 looks at alternatives to the dominant model: these involve identifying hubs and authorities for information, discovering web communities, and employing techniques of machine learning and user feedback; we also briefly describe the standard technique of bibliometric evaluation and relate it to prestige in the web. Finally—and most importantly—we

take a critical look at how people use search engines. Natural biases arrive which invariably remain hidden. Users, focused on their information retrieval tasks, do not reflect on the selection mechanism that serve information up to their desktop and dictate what they actually see. Questions naturally arise concerning privacy and the use of personal information. Finally, in Section 8, we reflect upon possible solutions to the problems raised.

2 Development of web search

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 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. We explain how this works below.

But first let us return to the classic model of full-text retrieval. Given a query and a set of documents, the task is to locate those documents that are most relevant to the query [Witten 99]. This problem was studied comprehensively and in great detail from the 1960s onwards. The web, when it came along, provided a massive, universally accessible set of documents. Computer scientists eagerly rushed off to apply information retrieval techniques in this new adventure playground. The web presented great challenges because of its size—for at the time it was not easy to get hold of massive quantities of electronic text. There was also the fun of downloading or "crawling" the web so that full-text indexes could be produced, itself an interesting problem. Soon the first search engines appeared. They were marvelous systems that faithfully located all web pages containing the keywords you specified, and presented them in order of relevance. It was a triumph of software engineering that such huge indexes could be built at all, let alone consulted by many users simultaneously.

Prior to the inception of search engines, the web was of limited practical interest. Of course, it was nice to be able to read what others had written, and follow their hyperlinks to further interesting material. But seeking out new information was like looking for needles in haystacks. Search engines changed all that. When academics and researchers learned about search engines (for this was before the web's discovery by commerce), they started to learn more about the web.

While most were playing in this new sandbox and marveling at its toys and how easily you could find them, a perspicacious few noticed problems. Web queries are very short: usually one or two words. If a particular word occurs at most once per document, standard ranking algorithms return documents in order of length, shortest first—because brevity enhances the apparent significance of each word. In an early example from Google's pioneers, the word *University* returned a long but haphazard list of pages, sorted in order of the term's occurrence frequency and the document's length [Brin 98]. Another problem arose when people began putting words into their pages specifically to get noticed. Traditional information retrieval takes documents at face value: it assumes that the words they contain are a fair representation of what they are about. But if you want to promote your wares, why not include additional

text intended solely to increase its visibility? Unlike paper documents, words can easily be hidden in web pages, visible to programs but not to human readers.

In real life we assess the quality of information by what others say about it, not by how it describes itself. Likewise, one can determine what web pages are about by looking at the clickable text of hyperlinks that point to them, called "anchor text". This insight suggests including the words on all links into a page in the full-text index entries for that page, as though the anchor text were actually present in the page itself. In fact, these words should be weighted more highly than those in the page because external opinions are usually more accurate. In a world tainted with deceit this substantially improves retrieval effectiveness.

3 Measuring prestige

Suppose we want to list the pages that match a query in order of their prestige, rather the density of occurrence of the search terms as traditional relevance ranking techniques do. Prestige is "high standing achieved through success or influence." A metric called PageRank, introduced by Google's founders and used in various guises by other search engines too, measures the standing of a web page [Brin 98]. The implicit thesis is that prestige is a good way to determine authority, defined as "an accepted source of expert information or advice."

In a networked community, people reward success with links. Page authors link to other pages because they find them useful and informative—they are successful web pages. If a host of people link to the same one, that indicates prestige. Figure 1 shows a tiny (fictitious) fraction of the web, including links between pages. Which ones do you think are most authoritative? Page F has five incoming links, so there's a good chance that this page is more authoritative than the others. B is second best, with four links.

Counting links is a crude measure. Some web pages have thousands of outgoing links whereas others have just one or two. Rarer links are more discriminating and should weigh more than others. In Figure 1 the many links emanating from page A indicate that A is a prolific linker, so each link carries less weight. From F's point of view, the links from D and E may be more valuable than the one from A.

There's another factor: a link is more valuable if it comes from a prestigious page. The link from B to F may be better than the others into F because B is a more prestigious page. At first sight this smacks of the "old school tie" network of political elite, which bestows a phony and incestuous kind of prestige. But here it's different: prestige is not an accident of breeding but must be earned by attracting links. Admittedly this factor involves a certain circularity, and without further analysis it's not clear that it can be made to work.

Underlying these ideas is the assumption that all links are bone fide ones. We fretted earlier that deceitful authors could insert misleading words into their pages to attract attention and ensure that they were returned more often as search results. Could they not also establish a fake kind of prestige by establishing phony links to their page? The answer is yes. But arranging phony links is not as easy as editing the page to include misleading words. What counts are links *in* to the page, not links from it to others. And placing thousands of links from another page does not help much

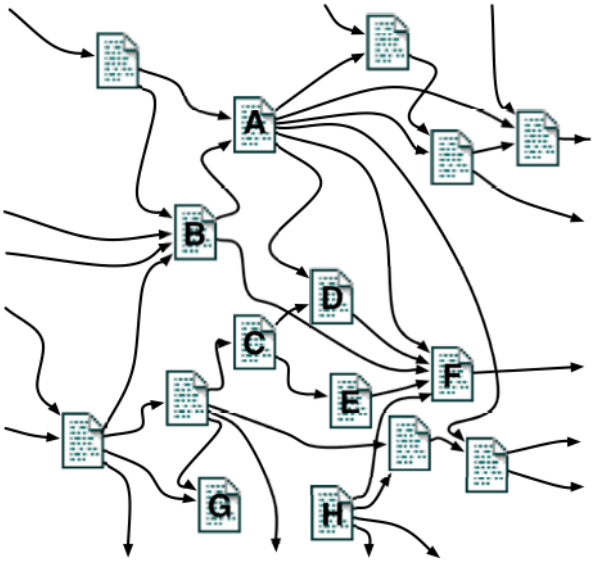

Figure 1: A tangled web

because inlinks are not just counted—their influence is attenuated by the host of outlinks from the linking page.

To summarize: the PageRank of a page is a number between 0 and 1 that measures its prestige. Each link into the page contributes to its PageRank. The amount it contributes is the PageRank of the linking page divided by the number of outlinks from it. The PageRank of any page is calculated by summing that quantity over all links into it. The value for D in Figure 1 is calculated by adding one-fifth of the value for A (because it has five outlinks) to one-half the value for C. Mathematically, the rank $r(q)$ of a page q is given by

$$r(q) = \sum_{\text{pages } p \text{ that link to } q} \frac{1}{o(p)} r(p)$$

where the sum is taken over all pages p that link to q, and $o(p)$ is the number of outlinks of page p.

3.1 Calculating PageRank

The definition is circular: how can you calculate the PageRank of a page without knowing the PageRanks of all the other pages? The answer is to use an iterative method. Start by randomly assigning an initial value to every page. The values could be chosen to be different, or they could all be the same: it doesn't matter (provided they're not all zero). Then recompute each page's rank by summing over its inlinks. If the initial values are thought of as an approximation to the true value of PageRank, then the new values are a better approximation. In successive iterations, the same method is used to recompute the rank for every page in the web. The process

terminates when, for every page, the next iteration turns out to give (almost) exactly the same rank as the previous one. With minor modifications discussed below (Section 4), this process is guaranteed to converge. The number of iterations depends on the desired accuracy (and other more technical factors). The problem can be formulated in terms of linear algebra, and the web presents the largest practical problem in linear algebra that has ever been contemplated.

The web's "connection matrix" is an array $C(p,q)$ whose rows represent the links out of a particular web page, and whose columns represent the links into another web page. $C(p,q)$ is 1 if page p contains a link to page q; otherwise it is 0. The dimensions of the array are the number of pages in the web: huge. Most entries are 0 since the probability of one randomly chosen web page linking to another is very small. Mathematically speaking,

$$r(q) = \sum_{\text{all web pages } p} r(p)\frac{C(p,q)}{o(p)}.$$

More compactly, using matrix notation,

$$\underline{r} = \underline{\underline{N}}\,\underline{r}$$

where $\underline{\underline{N}}$ is a normalized version of the web connection matrix in which each element $C(p,q)$ is divided by the number of outlinks from p. (In practice, for this to work $\underline{\underline{N}}$ must be an irreducible stochastic matrix, and some small modifications must be made to it to ensure this; we describe these below in Section 4.) This equation means that the vector $\underline{r}$ of page ranks is the principal eigenvector of $\underline{\underline{N}}$ (with eigenvalue 1). Such computations have been studied for 150 years and many methods have been developed for solving just the kind of problem that PageRank presents. The web presents the largest, most practical, matrix problem ever encountered. Suddenly the expertise of mathematicians is at the very core of companies that trade for billions of dollars on the stock exchange.

The iterative technique described above is what search engines use today, but the precise details are only known to insiders. The accuracy used for the final values probably lies between 10^{-9} and 10^{-12}. Brin and Page [Brin 98] reported 50 iterations for a much smaller version of the web than today's, before the details became commercial; several times as many iterations are probably needed now. Google is thought to run programs for several days to perform the PageRank calculation for the entire web, and the entire operation is—or at any rate used to be—performed every few weeks.

How should the initial values be set? The current values of PageRank would seem an excellent choice to begin the iteration; unfortunately this does not reduce the number of iterations significantly over a random starting-point. Some pages—for example, those concerning news and current events—need updating far more frequently than once every few weeks. Incrementally updating the PageRank calculation is an important practical problem: you somehow want to use the old values and take into account changes to the web—both new pages and new links on old pages. There are ways of doing this, but they don't apply on a sufficiently large scale. Approximate updating is an active research area that has received a great deal of attention in search engine companies. And none of it is published.

3.2 Combining prestige and relevance

To deal with one-word queries, all pages that contain the search term could be located and returned in order of prestige. But shouldn't the number of occurrences of the word be taken into account? And what about multiword queries? Ordinary ranked retrieval treats multiword queries as OR queries and calculates a relevance measure for each document to determine the order of results. The prestige model suggests combining this with PageRank.

Because the web is so vast, popular search engines treat all queries as AND queries, so that only pages that contain all the search terms are considered. However, there is still the question of how many times the search terms appear in each page. Moreover, search engines modulate the influence of terms in the page using heuristics. A word appearance is more important if it

- occurs in anchor text
- occurs in the title tag
- occurs in the document's URL
- occurs in an HTML heading
- occurs in capital letters
- occurs in a larger font than the rest of the document
- occurs early on in the document
- occurs in a HTML metatag.

A set of query terms is more influential if they

- appear close together
- appear in the right order
- appear as a phrase.

These could all affect the order in which search results are returned. However, the precise set of factors used by search engines today is unknown—and changes over time. For example, in an ideal world words in HTML metatags ought to be especially important, because these are intended to help characterize the content of the document. But tags are widely misused to give an erroneous impression of what the document was about. Today's search engines may treat them as more influential, or less, or ignore them, or even use them as negative evidence. Who knows? Only insiders.

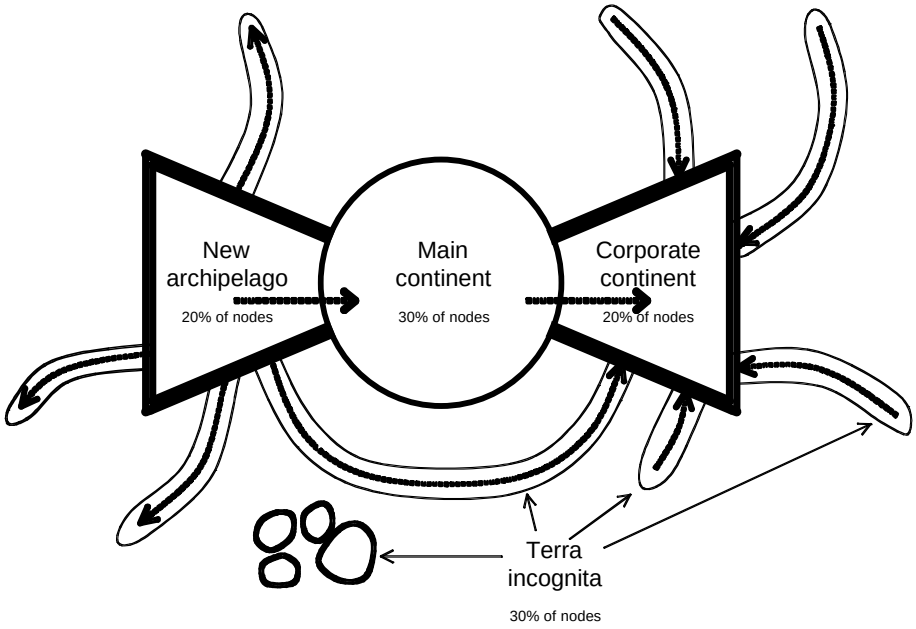

Figure 2: Chart of the web

All these factors are combined with the PageRank of the returned document—which clearly plays a dominant role—into a single measure, and this is used to sort documents for presentation to the user. The recipe is a closely guarded secret—think of it as the crown jewels of the search engine company. It changes from one month to the next to help fight spam.

4 The random surfer

Imagine a web surfer who chooses an outlink at random from the page he is on and follows it, continuing forever. The probability of taking any particular link is smaller if there are many outlinks, which is exactly the behavior we want from PageRank. It turns out that the PageRank of a given page is proportional to the probability that the random surfer lands on that page. But this model highlights two problems. The surfer flows through the tangled web of Figure 1, arriving at a page through its inlinks and leaving it through its outlinks. What if there are no inlinks (page H)—or no outlinks (G)?

Figure 2 shows a chart of the web, produced during a systematic study of its hyperlink structure in 1999 [Broder 00]. The shape is reminiscent of a bowtie. The central knot is a giant subnet that we call the *main continent*, a large strongly connected structure in which all pages are linked together (the continent metaphor was introduced in [Barabási 02]). Here you can travel from one page to any other by

following the links—just as when surfing with a browser. This is where most surfing takes place. Crawlers explore it all. It's the dominant part of the web.

The directed hyperlinks create other regions, also shown on the chart. There are three, each the same size or slightly smaller than the main continent. The *new archipelago*, though shown as a solid block, is in fact a large group of fragmented islands; the main continent can be reached from each island by following links. Most pages here are likely to be quite new, and haven't received many links. The archipelago's youth explains its fragmentation. The fact that it contains no large strongly connected components reflects its recent evolution—it's like the dawn of the web. It does nevertheless contain some old pages to which no-one in the main continent has thought fit to link.

The corporate continent comprises pages that are all reachable from the main continent. Some are sinks, pages without any outward links at all (for example, most Word and PDF files); however, these are a minority. This continent is highly fragmented and comprises an immense number of connected islands. Although it contains the second largest strongly connected component in the web (not shown on the chart), this component is hundreds of times smaller than the main continent. The corporate continent includes company websites that do not link out to the main continent. Though relatively small compared to the main continent, these islands are significantly larger than those of the new continent.

Terra incognita is the remainder of the universe. The distinctive feature of its pages is that surfers who reach them haven't come from the main continent and won't get there in the future—though they may have traveled from the new archipelago, and they may end up in the corporate continent. Linked trails in terra incognita do not cross the main part of the web: the pages are simply disconnected from it. However, as shown in the chart, pages in terra incognita can be connected to the new or corporate continents, giving rise to tendrils of two different kinds. Some tendrils only receive links from pages in the new continent, whereas others send links to the corporate continent. Like those in the new archipelago, pages in terra incognita are likely to be new.

Now the problem raised by a page with no outlinks (G) becomes apparent: it's a PageRank sink because once the surfer has entered he cannot leave. More generally, a set of pages might link to each other but not to anywhere else. This incestuous group is also a PageRank sink: the surfer gets stuck in a trap. He has reached an island that forms part of the corporate continent. As for a page with no inlinks (H), the surfer never reaches it. In fact, he never reaches any group of pages that has no inlinks from the rest of the web, even though it may have internal links, and outlinks to the web at large. He never visits the new archipelago.

Both these problems can be rectified by making small adjustments to the matrix $\underline{\underline{N}}$. A page with no outlinks is given a complete set of notional outlinks leading to every other page with a tiny probability. This makes the matrix "stochastic." And for pages with no inlinks, the matrix is adjusted to make the surfer, with a certain small probability, arrive at a randomly chosen web page instead of following a link from the one he is on. Then, if he's stuck in the corporate continent the surfer will eventually "teleport" out of it, and he will eventually explore the new archipelago by teleporting into it. This makes the matrix "irreducible." Mathematically

$$\underline{r} = \underline{\underline{N}}' \, \underline{r},$$

where $\underline{\underline{N}}'$ is a modified version of $\underline{\underline{N}}$ which is stochastic and irreducible.

The teleport probability has a strong influence on the rate of convergence of the iterative algorithm—and on the accuracy of its results. If it were set to 1 so that the surfer always teleported, the link structure of the web would have no effect on PageRank, and no iteration would be necessary. If it were 0 and the surfer never teleported, the calculation would not converge at all. Early published experiments used a teleportation probability of 0.15; some speculate that search engines increase it a little to hasten convergence.

Instead of teleporting to a randomly chosen page, you could choose a predetermined probability for each page, and—once you had decided to teleport—use that probability to determine where to land. This does not affect the calculation. But it does affect the result. If a page were discriminated against by receiving a smaller probability than the others, it would end up with a smaller PageRank than it deserves. This gives search engine operators an opportunity to influence the results of the calculation—an opportunity that they use to discriminate against certain sites (e.g., ones they believe are trying to gain an unfair advantage by exploiting the PageRank system).

5 Alternatives to PageRank

The success of the PageRank concept was responsible for elevating Google to the position of the world's preeminent search engine. However, it is not the only game in town.

5.1 Finding hubs and authorities for a query

At the same time as Brin and Page were developing PageRank, an alternative was being investigated by computer science academic Jon Kleinberg [Kleinberg 98]. Called HITS for Hypertext-Induced Topic Selection, it has the same self-referential character as PageRank but the details are intriguingly different. Though their work proceeded independently, Page/Brin and Kleinberg's 1998 papers cite each other.

Whereas PageRank is a single measure of the prestige of each page, HITS divides pages into two classes: hubs and authorities. A hub has many outlinks; a good hub contains a list of useful resources on a single topic. An authority has many inlinks, and comprises a useful source document. Good hubs point to good authorities; conversely, good authorities are pointed to by good hubs. In Figure 1, page F looks like a good authority while A looks like a good hub. Each page has two measures, hub score and authority score. The hub score is the sum of the authority scores of all the pages it links to. The authority score is the sum of the hub scores of all the pages that link to it. Like PageRank, the solution can be formulated as a problem in linear algebra based on the web's connection matrix.

Unlike PageRank, the HITS method does not apply this technique to the whole web once and for all. Instead hub and authority scores are computed that are particular to the query at hand. Given a query, a set of related pages is determined as follows.

First all pages that contain the query terms are located and placed in the set. Next, the pages that are linked to by these original set members, and the pages the original set members link to, are added to the set. The process could be continued, adding two-link neighbors, three-link neighbors, and so on, but in practice the set is probably quite large enough already. Indeed, it may be too large: perhaps having included all query-term pages only a few pages that each one points to should be added, along with a few pages that point to each query-term page.

The result is a subgraph of the web called the query's *neighborhood graph*. For each of its pages, hub and authority scores are derived using the above technique. In fact, division into hubs and authorities sidesteps some of PageRank's convergence problems. However, lesser convergence problems can still arise, and similar solutions apply. Of course, the computational load is far smaller than for PageRank because we are dealing with a tiny subset of the web. On the other hand the work must be redone for every query, rather than once a month as for PageRank.

A nice feature of HITS is that it helps with synonyms. A classic problem of information retrieval is that many queries do not return pertinent documents merely because they use slightly different terminology. For example, search engines like Google do not return a page unless it contains all the query terms. HITS includes in the neighborhood graph pages that are linked to by query-term pages, and pages that link to query-term pages. This greatly increases the chance of pulling in pages containing synonyms of the query terms. It also increases the chance of pulling in other relevant pages. For example, there is no reason to expect that Toyota or Honda's home pages should contain the term *automobile manufacturers*, yet they are very much authoritative pages on the subject.

5.2 Discovering web communities

Within the web it seems plausible that you many different communities could be identified. Each would contain many links to the web pages of other community members, along with a few external links. Scientific communities are a good example: biologists tend to link to web pages of other biologists. In fact, clusters of self-referential nodes in a graph can be identified using the matrix technique. The principal eigenvector represents the dominant web community, and other eigenvectors relate other communities. It would be fascinating to apply this method of social network analysis to the web graph, but because of its immense size this is impractical [Gibson 98].

The fact that HITS works with a far more manageable matrix than the entire web opens the door to new possibilities. The hub and authority scores found by the iterative algorithm we have described represent the structure of the dominant community. Other techniques can be used to calculate alternative sets of scores, many of which represent identifiable subcommunities of the original set of web pages. These might give more focused results for the user's query.

Consider the neighborhood graph for a query, and recall that pages containing the query terms have been augmented by adding linking and linked-to pages as well. These pages will likely cover several different communities that use the same terms in different ways. Each community has its own hubs and authorities. Rather than simply returning a list of search results, a query gives an entrée into the entire space, created on the fly by the search engine. The original broad topic is distilled into subtopics,

and high-quality pages are identified for each one. For example, a search for *soprano* might identify sets of web pages associated with *Marie-Adele McArthur* (a renowned New Zealand soprano), *Three Sopranos* (the operatic trio), *The Sopranos* (a megapopular U.S. television show) and several other choices.

Is it really feasible for a search engine to do all this work in response to a single user's query? Are social network algorithms efficient enough to operate on the fly? [Davison 99] report that it takes less than a minute of computation to process each query—fast, but nowhere near enough to satisfy impatient users, and far beyond what a mass-market search engine could afford. Of course, just how search engines work is a closely guarded secret. This is why mathematicians expert in matrix computation suddenly became hot property.

5.3 Learning to rank

Analyzing huge networks containing immense amounts of implicit information will remain a fertile research area for decades to come. Today it is bubbling with activity, as befits a business populated by a mixture of mathematicians and multimillionaires. We can be sure that radical new methods will appear, perhaps ousting today's megadragons. (Unfortunately, if we want to know how they work we will probably just have to guess.) The book is by no means closed on search engine technology.

Techniques of machine learning are being recruited to the task of ranking web pages. To do this, first create a "training set" with many examples of documents that contain the terms in a query, along with human judgments about how relevant they are. The learning algorithm analyzes this training data and comes up with a way to predict the relevance judgment for any document and query. This is used to rank queries in the future.

Machine learning involves straightforward algorithms that take a set of training data and produce a model for calculating judgments on new data [Witten 05]. A common method is to use numeric weights to combine different features of the document—for example, the features listed in Section 3.2 above. Each feature might simply be multiplied by its weight and summed together. The weights reflect the relative importance of the features, and training data is used to derive suitable values for them—values that approximate the relevance judgments assigned to the training examples by human evaluators. Of course, we want the system to work well not just for the training data, but for all other documents and queries too.

The system cannot come up with a different set of weights for each query—there are an infinite number of possible queries. Instead, for each document a set of feature values is calculated that depend on the query term—for example, how often it appears in anchor text, whether it occurs in the title tag, whether it occurs in the document's URL, how often it occurs in the document itself. And for multi-term queries, how often two different terms appear close together in the document, and so on. There are many possible features: typical algorithms for learning ranks use several hundred—let's say a thousand. Given a query, a thousand feature values are computed for every document that contains the query terms, and combined to yield its ranking. They may just be weighted and added together. Some machine learning techniques combine them in ways that are slightly more complex than this, but the principle remains the same.

The problem is to derive the one thousand weights from the training data in a way that yields a good approximation to actual human judgments. There are many different techniques of machine learning. Microsoft's MSN search engine uses a technique called RankNet that employs a "neural net" learning scheme [Burges 05]; a related algorithm is described by [Diligenti 03]. Despite its brainy anthropomorphic name, this need not be any more complex than calculating a weight for each feature and summing them up [Witten 05].

5.4 User feedback

Learning techniques have the potential to improve their performance as new data is gathered from users of a search engine. Judgments of which documents are approved and disapproved give a rich set of additional information for the search engine to use. In practice, harried users are hardly likely to give explicit feedback about every document in the search results. However, information can be gleaned from the user's subsequent behavior. Which documents does she click on, how long does she dwell on each one, and are her needs satisfied or does she return to search again? There is much information here, information that could be used to improve search performance in general as well as to improve results for specific searches. Typical queries are not unique, but are issued many times by different users. The subsequent behavior of each user could be integrated to provide an enormous volume of information about which documents best satisfy that particular query.

You might object that once you have made a query, search engines do not get to see your subsequent behavior. However, they can easily intercept the clicks you make on the search results page: they simply return information about which link is clicked by redirecting the link back through their own site. It does seem difficult to determine information about your subsequent behavior: how long you spend with each document, or what you do next. But that depends. Who wrote the web browser? Have you downloaded the Google toolbar (a browser add-on)? If the outcome improves the results of your searches, you might well be prepared to share this information with the dragons. After all, you're sharing your queries—which may reflect your most intimate hopes and fears.

Full-text search, the classic retrieval method, uses information supplied by authors of the document text. Link analysis uses information supplied by other authors—ones who link to the document. User feedback, the next wave, uses information supplied by users—readers, not writers. It's potentially far more powerful because it is these end users, not the authors, who are actually doing the searching.

5.5 Bibliometrics

The ideas underlying PageRank echo bibliometric techniques that are used to analyze the citation or cross-reference structure of the printed literature [Egghe 90]. Scientists are ranked on the basis of the citations that their papers attract. In general, the more citations your papers receive, the greater your prestige. But citations carry more weight if they come from someone who references selectively rather than citing copiously. There's also a parallel to teleporting: scientists with no citations at all still deserve a non-zero rank.

The *impact factor*, calculated each year by the Institute for Scientific Information (ISI) for a large set of scientific journals, is a widely used measure of a journal's importance [Garfield 72]. It has a huge, though controversial, influence on the way published research is perceived and evaluated. For a given journal in a given year it is defined as the average number of citations received by papers published in that journal over the previous two years, where citations are counted over all the journals that ISI tracks. In our terms, it is based on counting (rather than weighing) inlinks.

More subtle measures have been proposed, based on the observation that not all citations are equally important. Some have argued that a journal is influential if it is heavily cited by other influential journals—a circular definition just like the one we used earlier for "prestige." The connection strength from one journal to another is the fraction of the first journal's citations that go to papers in the second. In concrete terms, a journal's measure of standing, called its *influence weight*, is the sum of the influence weights of all journals that cite it, each one weighted by the connection strength. This is essentially the same as the recursive definition of PageRank (without the problem of pages with no inlinks or no outlinks). The random surfer model applies to influence weights too: starting with an arbitrary journal you choose a random reference appearing in it and move to the journal specified in the reference. A journal's influence weight is the proportion of time spent in that journal.

6 Making it work in practice

How search engines actually deliver the goods is one of the marvels of our world. Full-text search is an advanced technology. Although the concepts are simple, making them work is not. Leading search engines process millions of queries per second and respond to each one in half a second. They index many terabytes of text, and though disks may be large and cheap, organizing information on this scale is daunting. But the real problem is that with computers, time and space interact. For speed, everything must be in main memory. To scale up, everything must be on disk. Advanced computer science algorithms are required to manage the conflict. Couple full-text searching with link analysis: searching in a web. This involves advanced mathematics. Calculating PageRank, or building the neighborhood graph's connection matrix and analyzing it to determine communities, are not easy. Combining prestige and relevance, and optimizing the various factors involved, involves tedious experimentation with actual queries and painstaking evaluation of search results. All these affect your company's bottom line, in a hotly competitive market.

An even greater technical marvel than fast searching and web analysis is the standard of responsiveness and reliability set by search engines. Each search engine company has tens or hundreds of thousands of interlinked computers. In part, speed and reliability is obtained by having separate sites operate independently. If the lights go out in California the site in Texas is unaffected. Your query is automatically routed to one of these sites in a way that balances the load between them, and if one is down the others silently share its burden. At each site the index and the cached web pages are split into parts and shared between different computers—thousands of them. Vast arrays of standard machines are more cost effective than specially designed supercomputers. The components are like ordinary office workstations, loaded with as much memory and disk as they can hold without going to special hardware. But the

machines are not boxed like office workstations; they are naked and mounted *en masse* in custom-designed racks.

Creating and operating such network presents great challenges. Most search engine companies probably use the Linux operating system—Google certainly does. Wherever possible they use open source software—for databases, for compressing the text, for processing the images, for the countless routine tasks that must be undertaken to maintain the service to users. When you have ten thousand computers you must expect many failures. One calendar day times ten thousand corresponds to 30 years. With so many machines there will be hundreds of failures every day. The system monitors itself, notices wounds, isolates them, informs headquarters. Its human operators are kept busy swapping out complete units and replacing them with new ones. Since you must plan for failure anyway, why not buy cheaper, less reliable machines, with cheaper, less reliable memory?

7 How we use search engines

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 crosscheck the source of information. 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 remained unchanged over a period of 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.

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)

7.1 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 but 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.

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—also a controversial information source, and one that is potentially volatile. 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 site)
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 (Government tourism site)
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 (Government tourism site)	National Portal of India

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

7.2 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 behavior 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.

7.3 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. However, 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 favor 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?

7.4 Personal information

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 from anywhere in the world just by querying a search engine.

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 there's a 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, counselors, and tribunals that mediate our access to society too.

8 Towards solutions

In the previous section we 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 raised.

8.1 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 favor 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 far 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 judgments 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.

8.2 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 neighbors, 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.

8.3 Personalization

Personalization is perhaps the toughest question of all. Personalization of our searches for information, 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.

9 Conclusions

Full-text search is a remarkable technology that radically affects our society because it determines how most people, most of the time, locate information. We have explained how the major search engines work. But though the principles are known, the details are not. The reason is commercial: inner secrets are jealously guarded because web dragons compete with each other. But there's another, more sinister, factor: details *have to* be secret in order for the dragons to serve users well. Unfortunately, people and companies try to take advantage of them to promote their wares. To protect users against spamming—results that do not satisfy the user's information need but are artificially promoted for commercial reasons—the dragons *in principle* cannot publish their inner details.

Fundamental changes are in the air as the dragons flex their muscles and expand their empires. As communities become central to web search, life will grow tougher for spammers. Search engines are starting to acknowledge the importance of personal interaction; they are beginning to recognize communities too. Peer-to-peer technology will offer new perspectives in web search. Curated digital libraries may become trusted sources where we go to find truly authoritative information. While struggling for leadership, the dragons have started to broaden their mandate by supplying office tools that help users manage their own information. Our information environment is evolving in opposite directions: towards distributed desktop systems equipped with integrated search, and towards centralized data supported by global hosting services.

Which will win?—Only time will tell. But one thing is certain. The dragons are aiming higher than search. They will change the very way we work.

Acknowledgements

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.

References

[Barabási 02] Barabási, A.L. (2002) *Linked: the new science of networks*. Cambridge, Massachusetts: Perseus.

[Brin 98] Brin, S. and Page, L. (1998) "The anatomy of a large-scale hypertextual Web search engine." *Computer Networks and ISDN Systems*, Vol. 33, pp. 107–117.

[Broder 00] Broder, A.Z., Kumar, R., Maghoul, F., Raghavan, P., Rajagopalan, V., Stata, S., Tomkins, A. and Wiener, J.L. (2000) "Graph structure in the Web." *Computer Networks*, Vol. 33, pp. 309–320.

[Burges 05] Burges, C., Shaked, T., Renshaw, E., Lazier, A., Deeds, M., Hamilton, N. and Hullender, G. (2005) "Learning to rank using gradient descent." *Proc Int Conf on Machine Learning*, Bonn.

[Davison 99] Davison, B.D., Gerasoulis, A., Kleisouris, K., Lu, Y., Seo, H., Wang, W. and Wu, B. (1999) "DiscoWeb: Applying link analysis to Web search." (Extended abstract) *Proc Int World Wide Web Conf*, pp. 148–149. ACM Press, New York.

[Diligenti 03] Diligenti, M., Maggini, M. and Gori, M. (2003) "A learning algorithm for web page scoring systems." *Proc. Int Joint Conf on Artificial Intelligence*, pp. 575-580. Acapulco, Mexico.

[Egghe 90] Egghe, L. and Rousseau, R. (1990) *Introduction to Informetrics*. Elsevier.

[Garfield 72] Garfield, E. (1972) "Citation analysis as a tool in journal evaluation," *Science*, 178, pp. 471–479.

[Gibson 98] Gibson, D., Kleinberg, J. and Raghavan, P. (1998) "Inferring Web communities from link topology." *Hypertext 1998*, pp. 225–234.

[Kleinberg 98] Kleinberg, J. (1998) "Authoritative sources in a hyperlinked environment." *Proc ACM-SIAM Symposium on Discrete Algorithms*. Extended version in *J ACM*, Vol. 46 (1999), pp. 604–632.

[Witten 99] Witten, I.H., Moffat, A. and Bell, T.C. (1999) *Managing gigabytes: Compressing and indexing documents and images*. Morgan Kaufmann, San Francisco.

[Witten 05] Witten, I.H. and Frank, E. (2005) *Data mining*. Morgan Kaufmann, San Francisco, second edition.

[Witten 07] Witten, I.H., Gori, M. and Numerico, T. (2007) *Web dragons: inside the myths of search engine technology*. Morgan Kaufmann, San Francisco.