Hypersearching the Web

With the volume of on-line information in cyberspace growing at a breakneck pace, more effective search tools are desperately needed. A new technique analyzes how Web pages are linked together

by Members of the *Clever* Project

Every day the World Wide Web grows by roughly a million electronic pages, adding to the hundreds of millions already on-line. This staggering volume of information is loosely held together by more than a billion annotated connections, called hyperlinks. For the first time in history, millions of people have virtually instant access from their homes and offices to the creative output of a significant—and growing—fraction of the planet's population.

But because of the Web's rapid, chaotic growth, the resulting network of information lacks organization and structure. In fact, the Web has evolved into a global mess of previously unimagined proportions. Web pages can be written in any language, dialect or style by individuals with any background, education, culture, interest and motivation. Each page might range from a few characters to a few hundred thousand, containing truth, falsehood, wisdom, propaganda or sheer nonsense. How, then, can one extract from this digital morass high-quality, relevant pages in response to a specific need for certain information?

In the past, people have relied on search engines that hunt for specific words or terms. But such text searches frequently retrieve tens of thousands of pages, many of them useless. How can people quickly locate only the information they need and trust that it is authentic and reliable?

We have developed a new kind of search engine that exploits one of the Web's most valuable resources—its myriad hyperlinks. By analyzing these interconnections, our system automatically locates two types of pages: authorities and hubs. The former are deemed to be the best sources of information on a particular topic; the latter are collections of links to those locations. Our methodology should enable users to locate much of the information they desire quickly and efficiently.

The Challenges of Search Engines

Computer disks have become increasingly inexpensive, enabling the storage of a large portion of the Web at a single site. At its most basic level, a search engine maintains a list, for every word, of all known Web pages containing that word. Such a collection of lists is known as an index. So if people are interested in learning about acupuncture, they can access the "acupuncture" list to find all Web pages containing that word.

Creating and maintaining this index is highly challenging [see "Searching the Internet," by Clifford Lynch; Scientific American, March 1997], and determining what information to return in response to user requests remains daunting. Consider the unambiguous query for information on "Nepal Airways," the airline company. Of the roughly 100 (at the time of this writing) Web pages containing the phrase, how does a search engine decide which 20 or so are the best? One difficulty is that there is no exact and mathematically precise measure of "best"; indeed, it lies in the eye of the beholder.

Search engines such as AltaVista, Infoseek, HotBot, Lycos and Excite use heuristics to determine the way in which to order—and thereby prioritize—pages. These rules of thumb are collectively known as a ranking function, which must apply not only to relatively specific and straightforward queries ("Nepal Airways") but also to much more general requests, such as for "aircraft," a word that appears in more than a million Web pages. How should a search engine choose just 20 from such a staggering number?

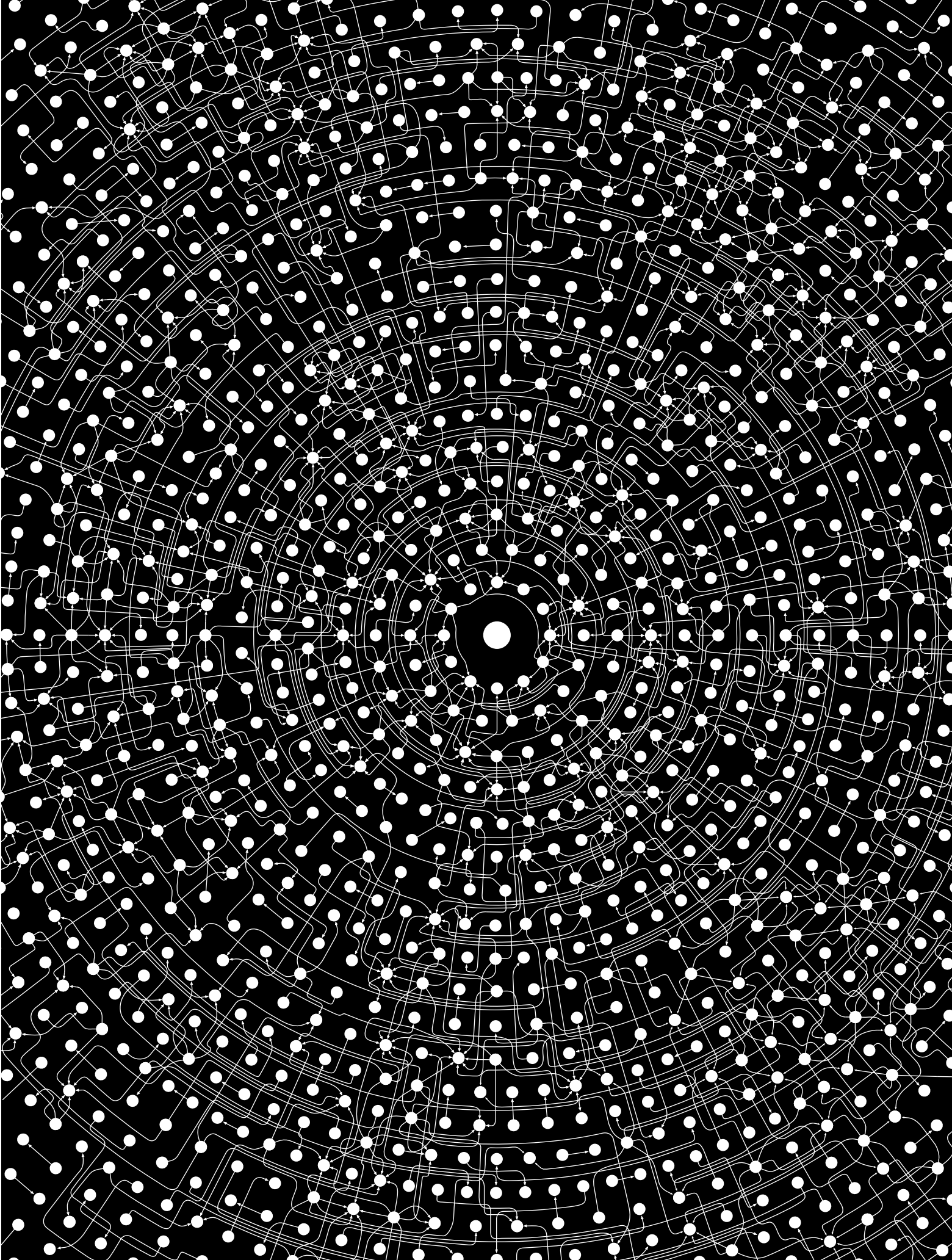
Simple heuristics might rank pages by the number of times they contain the query term, or they may favor instances in which that text appears earlier. But such approaches can sometimes fail spectacularly. Tom Wolfe's book *The Kandy-Kolored Tangerine-Flake Streamline Baby* would, if ranked by such heuristics, be deemed very relevant to the query "hernia," because it begins by repeating that word dozens of times. Numerous extensions to these rules of thumb abound, including approaches that give more weight to words that appear in titles, in section headings or in a larger font.

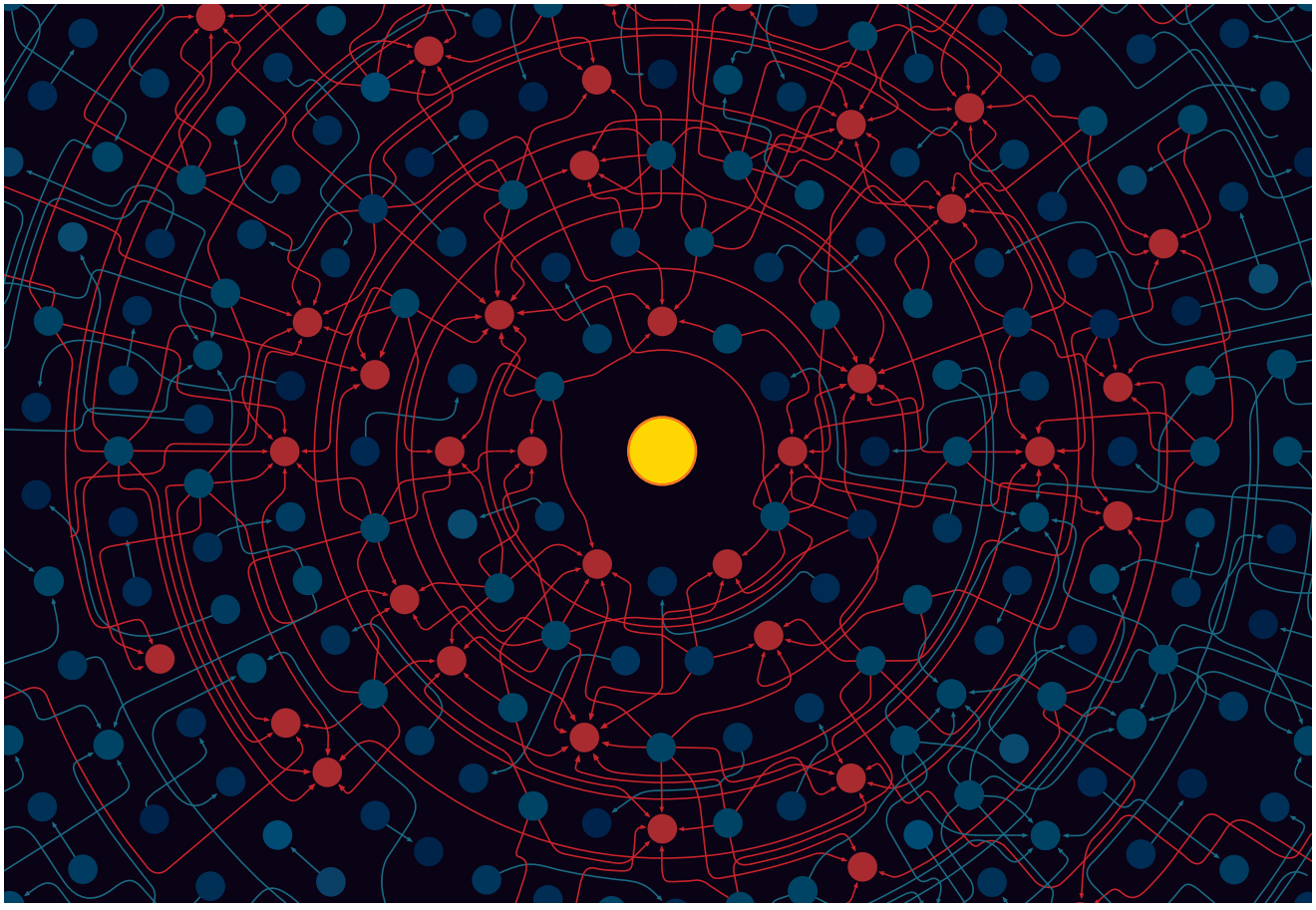
Such strategies are routinely thwarted by many commercial Web sites that design their pages in certain ways specifically to elicit favorable rankings. Thus, one encounters pages whose titles are "cheap airfares cheap airfares cheap airfares." Some sites write other carefully chosen phrases many times over in colors and fonts that are invisible to human viewers. This practice, called spamming, has become one of the main reasons why it is currently so difficult to maintain an effective search engine.

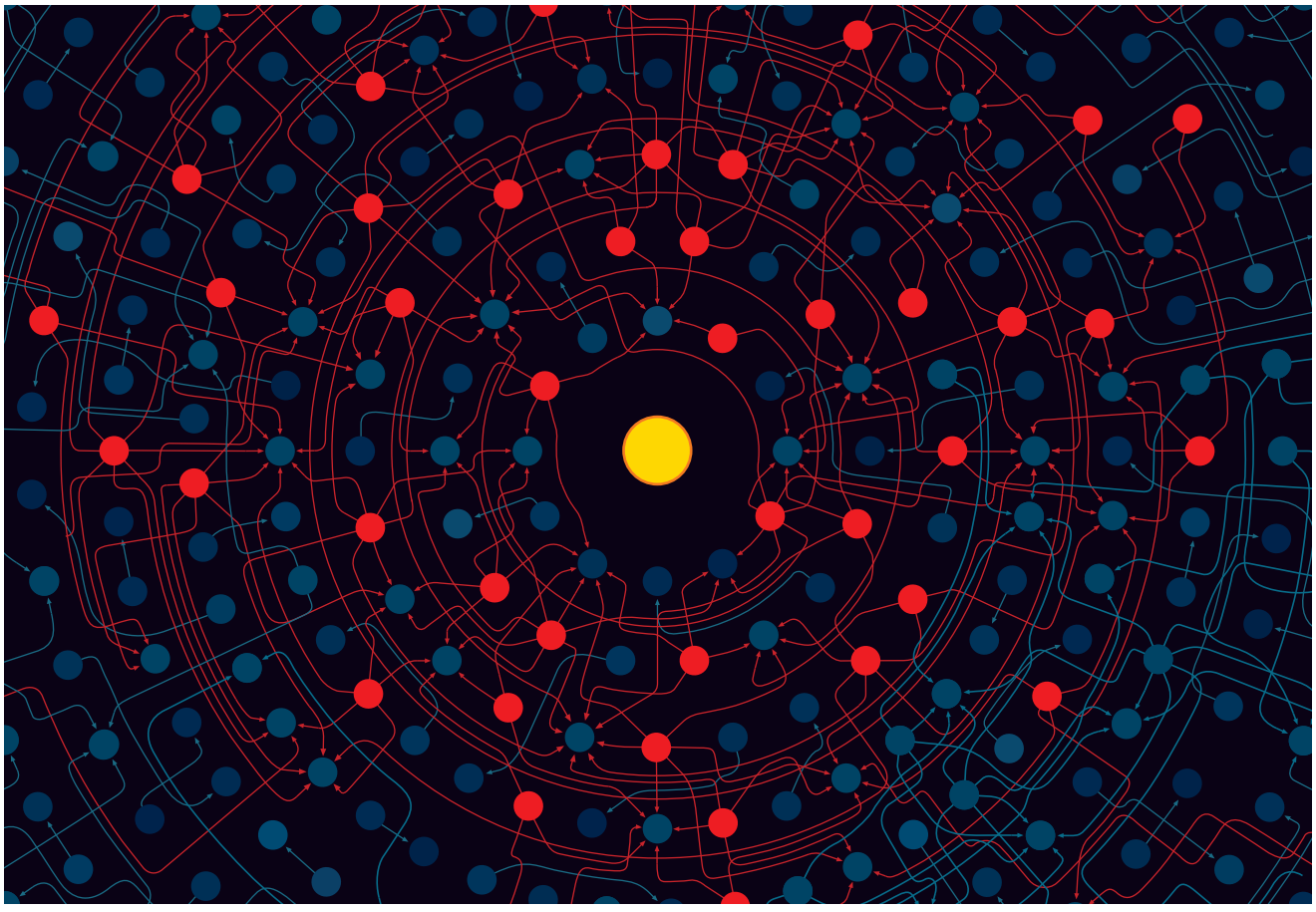
Spamming aside, even the basic assumptions of conventional text searches

ALL ILLUSTRATIONS BY BRYAN CHRISTIE

WEB PAGES (*white dots*) are scattered over the Internet with little structure, making it difficult for a person in the center of this electronic clutter to find only the information desired. Although this diagram shows just hundreds of pages, the World Wide Web currently contains more than 300 million of them. Nevertheless, an analysis of the way in which certain pages are linked to one another can reveal a hidden order.





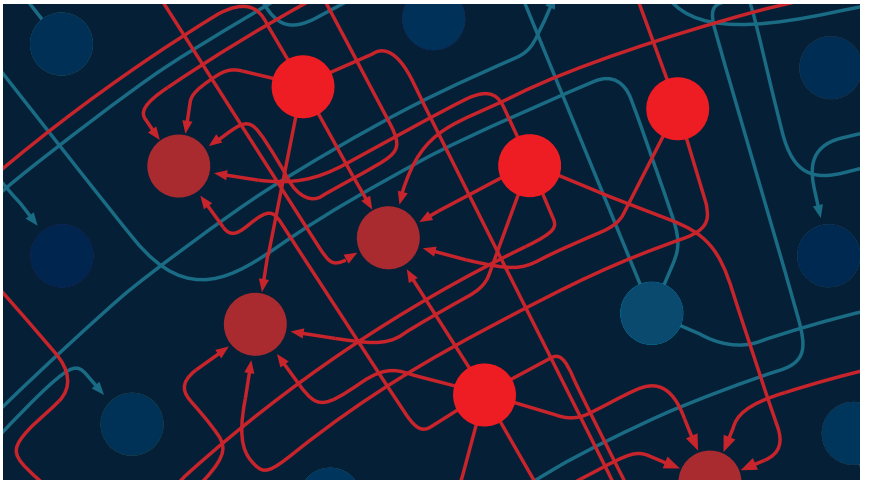


are suspect. To wit, pages that are highly relevant will not always contain the query term, and others that do may be worthless. A major cause of this problem is that human language, in all its richness, is awash in synonymy (different words having the same meaning) and polysemy (the same word having multiple meanings). Because of the former, a query for "automobile" will miss a deluge of pages that lack that word but instead contain "car." The latter manifests itself in a simple query for "jaguar," which will retrieve thousands of pages about the automobile, the jungle cat and the National Football League team, among other topics.

One corrective strategy is to augment search techniques with stored information about semantic relations between words. Such compilations, typically constructed by a team of linguists, are sometimes known as semantic networks, following the seminal work on the WordNet project by George A. Miller and his colleagues at Princeton University. An index-based engine with access to a semantic network could, on receiving the query for "automobile," first determine that "car" is equivalent and then retrieve all Web pages containing either word. But this process is a double-edged sword: it helps with synonymy but can aggravate polysemy.

Even as a cure for synonymy, the solution is problematic. Constructing and maintaining a semantic network that is exhaustive and cross-cultural (after all, the Web knows no geographical boundaries) are formidable tasks. The process is especially difficult on the Internet, where a whole new language is evolving—words such as "FAQs," "zines" and "bots" have emerged, whereas other words such as "surf" and "browse" have taken on additional meanings.

Our work on the Clever project at IBM originated amid this perplexing array of issues. Early on, we realized that the current scheme of indexing and retrieving a page based solely on the text it contained ignores more than a billion carefully placed hyperlinks that reveal the relations between pages. But how exactly should this information be used?



FINDING authorities and hubs can be tricky because of the circular way in which they are defined: an authority is a page that is pointed to by many hubs; a hub is a site that links to many authorities. The process, however, can be performed mathematically. Clever, a prototype search engine, assigns initial scores to candidate Web pages on a particular topic. Clever then revises those numbers in repeated series of calculations, with each iteration dependent on the values of the previous round. The computations continue until the scores eventually settle on their final values, which can then be used to determine the best authorities and hubs.

When people perform a search for "Harvard," many of them want to learn more about the Ivy League school. But more than a million locations contain "Harvard," and the university's home page is not the one that uses it the most frequently, the earliest or in any other way deemed especially significant by traditional ranking functions. No entirely internal feature of that home page truly seems to reveal its importance.

Indeed, people design Web pages with all kinds of objectives in mind. For instance, large corporations want their sites to convey a certain feel and project a specific image—goals that might be very different from that of describing what the company does. Thus, IBM's home page does not contain the word "computer." For these types of situations, conventional search techniques are doomed from the start.

To address such concerns, human architects of search engines have been tempted to intervene. After all, they believe they know what the appropriate responses to certain queries should be, and developing a ranking function that will automatically produce those results has been a troublesome undertaking. So they could maintain a list of queries like "Harvard" for which they will override the judgment of the search engine with predetermined "right" answers.

This approach is being taken by a number of search engines. In fact, a service such as Yahoo! contains only human-selected pages. But there are countless possible queries. How, with a limited number of human experts, can one maintain all these lists of precomputed responses, keeping them reasonably complete and up-to-date, as the Web meanwhile grows by a million pages a day?

Searching with Hyperlinks

In our work, we have been attacking the problem in a different way. We have developed an automatic technique for finding the most central, authoritative sites on broad search topics by making use of hyperlinks, one of the Web's most precious resources. It is the hyperlinks, after all, that pull together the hundreds of millions of pages into a web of knowledge. It is through these connections that users browse, serendipitously discovering valuable information through the pointers and recommendations of people they have never met.

The underlying assumption of our approach views each link as an implicit endorsement of the location to which it

AUTHORITIES AND HUBS help to organize information on the Web, however informally and inadvertently. Authorities (●) are sites that other Web pages happen to link to frequently on a particular topic. For the subject of human rights, for instance, the home page of Amnesty International might be one such location. Hubs (●) are sites that tend to cite many of those authorities, perhaps in a resource list or in a "My Favorite Links" section on a personal home page.

points. Consider the Web site of a human-rights activist that directs people to the home page of Amnesty International. In this case, the reference clearly signifies approval.

Of course, a link may also exist purely for navigational purposes ("Click here to return to the main menu"), as a paid advertisement ("The vacation of your dreams is only a click away") or as a stamp of disapproval ("Surf to this site to see what this fool says"). We believe, however, that in aggregate—that is, when a large enough number is considered—Web links do confer authority.

In addition to expert sites that have garnered many recommendations, the Web is full of another type of page: hubs that link to those prestigious locations, tacitly radiating influence outward to them. Hubs appear in guises ranging from professionally assembled lists on commercial sites to inventories of "My Favorite Links" on personal home pages. So even if we find it difficult to define "authorities" and "hubs" in isolation, we can state this much: a respected authority is a page that is referred to by many good hubs; a useful hub is a location that points to many valuable authorities.

These definitions look hopelessly circular. How could they possibly lead to a computational method of identifying both authorities and hubs? Thinking of the problem intuitively, we devised the following algorithm. To start off, we look at a set of candidate pages about a particular topic, and for each one we make our best guess about how good a hub it is and how good an authority it is. We then use these initial estimates to jump-start a two-step iterative process.

First, we use the current guesses about the authorities to improve the estimates of hubs—we locate all the best authorities, see which pages point to them and call those locations good hubs. Second, we take the updated hub information to refine our guesses about the authorities—we determine where the best hubs point most heavily and call these the good authorities. Repeating these steps several times fine-tunes the results.

We have implemented this algorithm in Clever, a prototype search engine. For any query of a topic—say, acupuncture—Clever first obtains a list of 200 pages from a standard text index such as AltaVista. The system then augments these by adding all pages that link to and from that 200. In our experience, the resulting collection, called the root set, will typically contain between 1,000 and 5,000 pages.

For each of these, Clever assigns initial numerical hub and authority scores. The system then refines the values: the authority score of each page is updated to be the sum of the hub scores of other locations that point to it; a hub score is revised to be the sum of the authority scores of locations to which a page points. In other words, a page that has many high-scoring hubs pointing to it earns a higher authority score; a location that points to many high-scoring authorities garners a higher hub score. Clever repeats these calculations until the scores have more or less settled on their final values, from which the best authorities and hubs can be determined. (Note that the computations do not preclude a particular page from achieving a top rank in both categories, as sometimes occurs.)

The algorithm might best be understood in visual terms. Picture the Web as a vast network of innumerable sites, all interconnected in a seemingly random fashion. For a given set of pages containing a certain word or term, Clever zeroes in on the densest pattern of links between those pages.

As it turns out, the iterative summation of hub and authority scores can be analyzed with stringent mathematics. Using linear algebra, we can represent the process as the repeated multiplication of a vector (specifically, a row of numbers representing the hub or authority scores) by a matrix (a two-dimensional array of numbers representing the hyperlink structure of the root set). The final results of the process are hub and authority vectors that have equilibrated to certain numbers—values that reveal which pages are the best hubs and authorities, respectively. (In the world of linear algebra, such a stabilized row of numbers is called an eigenvector; it can be thought of as the solution to a system of equations defined by the matrix.)

With further linear algebraic analysis, we have shown that the iterative process will rapidly settle to a relatively steady set of hub and authority scores. For our purposes, a root set of 3,000 pages requires about five rounds of calculations. Furthermore, the results are generally independent of the initial estimates of scores used to start the process. The method will work even if the values are all initially set to be equal to 1. So the final hub and authority scores are intrinsic to the collection of pages in the root set.

A useful by-product of Clever's iterative processing is that the algorithm naturally separates Web sites into clusters. A search for information on abortion, for example, results in two types of locations, pro-life and pro-choice, because pages from one group are more likely to link to one another than to those from the other community.

From a larger perspective, Clever's algorithm reveals the underlying structure of the World Wide Web. Although the Internet has grown in a hectic, willy-nilly fashion, it does indeed have an inherent—albeit inchoate—order based on how pages are linked.

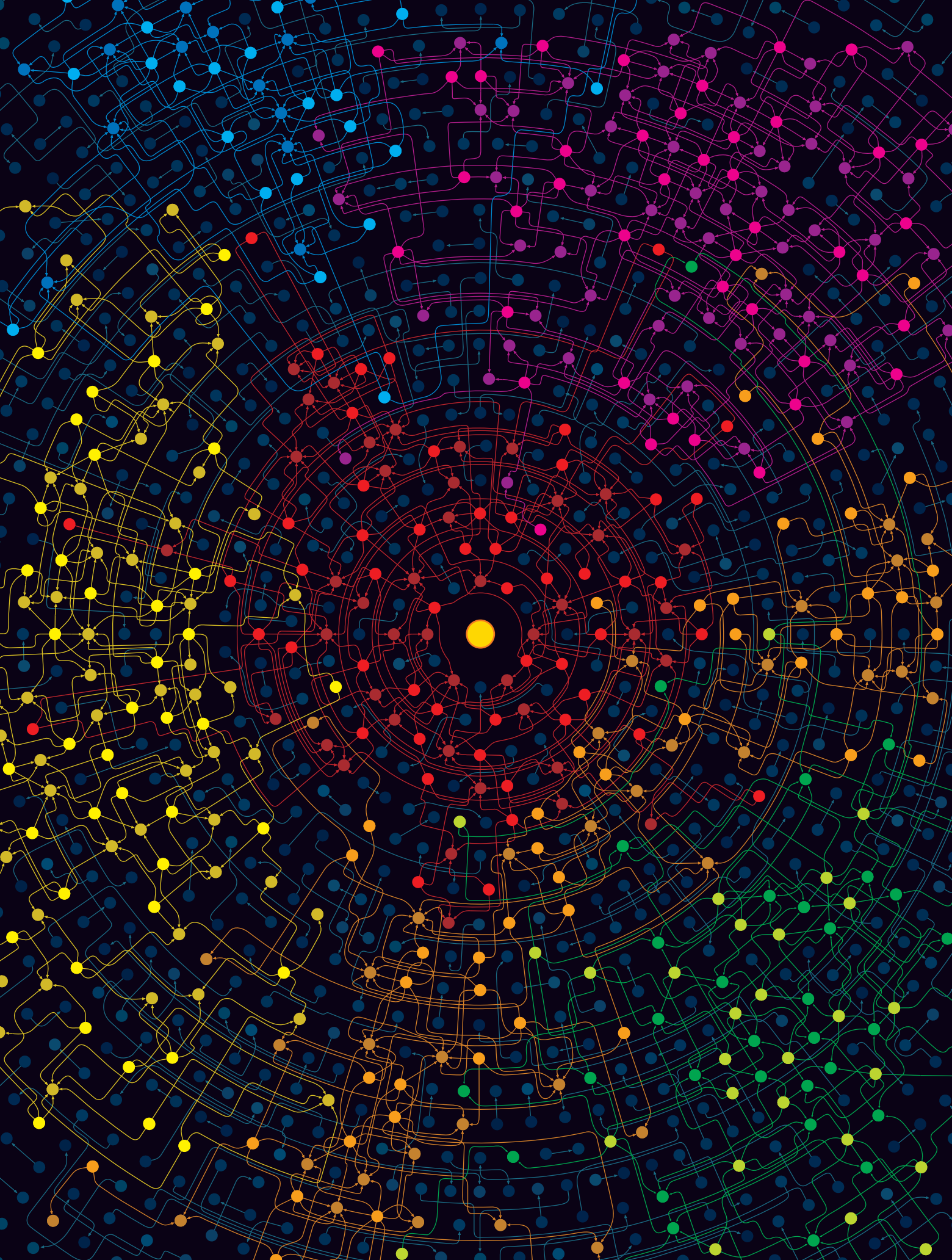
The Link to Citation Analysis

Methodologically, the Clever algorithm has close ties to citation analysis, the study of patterns of how scientific papers make reference to one another. Perhaps the field's best-known measure of a journal's importance is the "impact factor." Developed by Eugene Garfield, a noted information scientist and founder of *Science Citation Index,* the metric essentially judges a publication by the number of citations it receives.

On the Web, the impact factor would correspond to the ranking of a page simply by a tally of the number of links that point to it. But this approach is typically not appropriate, because it can favor universally popular locations, such as the home page of the *New York Times,* regardless of the specific query topic.

Even in the area of citation analysis, researchers have attempted to improve Garfield's measure, which counts each reference equally. Would not a better strategy give additional weight to citations from a journal deemed more important? Of course, the difficulty with this approach is that it leads to a circular definition of "importance," similar to the problem we encountered in specifying hubs and authorities. As early as

CYBERCOMMUNITIES (*shown in different colors*) populate the Web. An exploration of this phenomenon has uncovered various groups on topics as arcane as oil spills off the coast of Japan, fire brigades in Australia and resources for Turks living in the U.S. The Web is filled with hundreds of thousands of such finely focused communities.



1976 Gabriel Pinski and Francis Narin of CHI Research in Haddon Heights, N.J., overcame this hurdle by developing an iterated method for computing a stable set of adjusted scores, which they termed influence weights. In contrast to our work, Pinski and Narin did not invoke a distinction between authorities and hubs. Their method essentially passes weight directly from one good authority to another.

This difference raises a fundamental point about the Web versus traditional printed scientific literature. In cyberspace, competing authorities (for example, Netscape and Microsoft on the topic of browsers) frequently do not acknowledge one another's existence, so they can be connected only by an intermediate layer of hubs. Rival prominent scientific journals, on the other hand, typically do a fair amount of cross-citation, making the role of hubs much less crucial.

A number of groups are also investigating the power of hyperlinks for searching the Web. Sergey Brin and Lawrence Page of Stanford University, for instance, have developed a search engine dubbed Google that implements a link-based ranking measure related to the influence weights of Pinski and Narin. The Stanford scientists base their approach on a model of a Web surfer who follows links and makes occasional haphazard jumps, arriving at certain places more frequently than others. Thus, Google finds a single type of universally important page—intuitively, locations that are heavily visited in a random traversal of the Web's link structure. In practice, for each Web page Google basically sums the scores of other locations pointing to it. So, when presented with a specific query, Google can respond by quickly retrieving all pages containing the search text and listing them according to their preordained ranks.

Google and Clever have two main differences. First, the former assigns initial rankings and retains them independently of any queries, whereas the latter assembles a different root set for each search term and then prioritizes those pages in the context of that particular query. Consequently, Google's approach enables faster response. Second, Google's basic philosophy is to look only in the forward direction, from link to link. In contrast, Clever also looks backward from an authoritative page to see what locations are pointing there. In this sense, Clever takes advantage of the sociological phenomenon that humans are innately motivated to create hublike content expressing their expertise on specific topics.

The Search Continues

We are exploring a number of ways to enhance Clever. A fundamental direction in our overall approach is the integration of text and hyperlinks. One strategy is to view certain links as carrying more weight than others, based on the relevance of the text in the referring Web location. Specifically, we can analyze the contents of the pages in the root set for the occurrences and relative positions of the query topic and use this information to assign numerical weights to some of the connections between those pages. If the query text appeared frequently and close to a link, for instance, the corresponding weight would be increased.

Our preliminary experiments suggest that this refinement substantially increases the focus of the search results. (A shortcoming of Clever has been that for a narrow topic, such as Frank Lloyd Wright's house Fallingwater, the system sometimes broadens its search and retrieves information on a general subject, such as American architecture.) We are investigating other improvements, and given the many styles of authorship on the Web, the weighting of links might incorporate page content in a variety of ways.

We have also begun to construct lists of Web resources, similar to the guides put together manually by employees of companies such as Yahoo! and Infoseek. Our early results indicate that automatically compiled lists can be competitive with handcrafted ones. Furthermore, through this work we have found that the Web teems with tightly knit groups of people, many with offbeat common interests (such as weekend sumo enthusiasts who don bulky plastic outfits and wrestle each other for fun), and we are currently investigating efficient and automatic methods for uncovering these hidden communities.

The World Wide Web of today is dramatically different from that of just five years ago. Predicting what it will be like in another five years seems futile. Will even the basic act of indexing the Web soon become infeasible? And if so, will our notion of searching the Web undergo fundamental changes? For now, the one thing we feel certain in saying is that the Web's relentless growth will continue to generate computational challenges for wading through the ever increasing volume of on-line information. SA

The Authors

THE *CLEVER* PROJECT: Soumen Chakrabarti, Byron Dom, S. Ravi Kumar, Prabhakar Raghavan, Sridhar Rajagopalan and Andrew Tomkins are research staff members at the IBM Almaden Research Center in San Jose, Calif. Jon M. Kleinberg is an assistant professor in the computer science department at Cornell University. David Gibson is completing his Ph.D. at the computer science division at the University of California, Berkeley.

The authors began their quest for exploiting the hyperlink structure of the World Wide Web three years ago, when they first sought to develop improved techniques for finding information in the clutter of cyberspace. Their work originated with the following question: If computation were not a bottleneck, what would be the most effective search algorithm? In other words, could they build a better search engine if the processing didn't have to be instantaneous? The result was the algorithm described in this article. Recently the research team has been investigating the Web phenomenon of cybercommunities.

Further Reading

Search Engine Watch (www.searchenginewatch.com) contains information on the latest progress in search engines. The WordNet project is described in *WordNet: An Electronic Lexical Database* (MIT Press, 1998), edited by Christiane Fellbaum. The iterative method for determining hubs and authorities first appeared in Jon M. Kleinberg's paper "Authoritative Sources in a Hyperlinked Environment" in *Proceedings of the 9th ACM-SIAM Symposium on Discrete Algorithms,* edited by Howard Karloff (SIAM/ACM–SIGACT, 1998). Improvements to the algorithm are described at the Web site of the IBM Almaden Research Center (www.almaden.ibm.com/cs/k53/clever.html). *Introduction to Informetrics* (Elsevier Science Publishers, 1990), by Leo Egghe and Ronald Rousseau, provides a good overview of citation analysis. Information on the Google project at Stanford University can be obtained from www.google.com on the World Wide Web.