Hal Berghel

Cyberspace 2000: Dealing with Information Overload

ROBERT NEUBECKER

Francis Bacon is reported to have said that the three things that made his world different from that of the ancient Greeks and Romans were the printing press, the compass and gunpowder. It is instructive to note he didn't mention the water pump, the rigid horse collar, or lateen sails—all of which were critical to the advancement of agriculture and commerce.

One suspects Bacon was not interested in great technological tours de force per se, but in those technological advances that also stretched, or even tore, our social fabric and which irreversibly changed the way we looked at the world and each other. It was not enough to change the way we lived. To make Bacon's short list, a technology had to change the way we looked at life. Even the abilities to irrigate land and make it fertile, and navigate a ship into the wind as well as away from it, did not qualify. Bacon was looking for things as important to the 16th century as systems of writing were for the ancient peoples and stone tools and controlled fire were for the pre-historic.

We are in the midst of a technological revolution that will dramatically set our century apart from Bacon's—the digital networks and cyberspace. Together with perhaps fossil-fueled transportation, electricity, and television, cyberspace seems to most satisfy Bacon's requirement that a truly differentiating technology have far-reaching consequences for society. Of these four technologies, cyberspace is the only one that will come to be associated with the 21st century.

This installment of "Digital Village" is the first of several columns that will look at the future of cyberspace in the next decade—cyberspace in 2000. We'll attempt to foresee some of the important social issues, predict technological trends, and investigate promising, new, emerging technologies. We'll even offer some modest speculation, more than idly if not with perfect insight.

We begin with the challenge of information overload on cyberspace.

Internet Credibility

A point sometimes overlooked amidst all of the hoopla over the Internet is that it is now, and

will forever remain, credibility- and value-neutral. By this I mean the mere fact that a resource is available on the Internet does not provide any guarantee of importance, accuracy, utility or value. To be sure, specific Internet resources may develop credibility and value over time, but an Internet association as such and in general will never count for much in the way of credibility.

Failure to understand this has led to the proliferation of millions of individual and organizational vanity home pages and document clusters—frequently without any consideration given to potential use—as individuals and groups attempt to attach themselves to the trendy technology.

The most immediate cause of information overload on the Web is caused by the Web trying to fill the dual role of being both a private and public information and communication medium. Issues that are privately important tend to be publicly uninteresting. When the background noise of the medium drowns out most of the useful content for the wider audience, as is now happening on the Web, the effectiveness of the medium is undercut.

This, incidentally, is the same problem that ultimately ruined the citizen's band radio industry. The CB became a relatively useless communication medium because the industry did not anticipate, and could not handle, concurrent escalations in both volume of traffic and the proportion of noise. Fortunately, such propensity for self-destruction may be avoided in cyberspace because of its digital nature. Computational relief is forthcoming from all quarters.

Search Engines

The first attempt to deal with the information overload on the Web was the search engine. Modern, powerful search engines such as HotBot (www.hotbot.com) and Excite (www.excite.com) provide an example of how digital technology can be used to retrieve information. Well over 100 such search engines have been identified (128 are listed at ugweb.cs.ualberta.ca/~mentor02/search/search-all.html), each with its own particular search characteristics. In addition, meta-level search engines have also been developed, which utilize several object-level search engines in their operation, and then integrate their results. All4one (www.all4one.com) integrates four such search engines, Highway61 (www.highway61.com) seven, and SuperSeek (w3.superseek.com/superseek/) 10, to name a few (see Figure 1). Meta-level document indexers such as Yahoo (www.yahoo.com) and Galaxy (galaxy.einet.net/galaxy.html) are also available. Special purpose indexers also exist for such things as personal home pages, particular multimedia document types, academic interests, chemical structures, help-wanted ads, and so forth.

Figure 1. The SuperSeek meta search engine. Ten search engines are exploited in the meta-level search. SuperSeek uses Netscape frames to achieve the multi-paning interface for quick information uptake.

The present species of search engine typically consists of an HTML, form-based interface for submitting a query, an indexed database with an internal string matching routine, and some form of network indexer (which subsumes such entities as spiders, wanderers, crawlers, worms, ants and robots) that visits a specific set of network servers periodically and returns file or document-specific data for inclusion in the indexed database. Although the number and frequency of servers visited, the nature of the document extraction routines, and the robustness of the keyword-based, Boolean query interface varies by developer, all current search engines seem to target about the same level of search granularity.

Search engines as they now exist represent a primitive, first cut at efficient information access on the Internet. But the reality is

these search engines, which help to make cyberspace manageable, now index more chaff than wheat. As a data point, it is easy to verify that a search for a common technical term like "digital network" on a good search engine could produce hundreds of thousands of hits with very low overall yield in terms of useful information. There are, to be sure, seminal Web documents on digital networks, but finding them amidst pages and pages of content-free screen gumbo is akin to finding the proverbial needle in the haystack. This fact ensures the bright future of rival technologies.

The Fundamental Inadequacy of Search Engines

As an illustration of the magnitude of cyberspace, today's larger search engines now boast indices spanning over 50 million URLs, a relatively small percentage of which are likely to be of either immediate or enduring value to the broad Internet community. Whether the indexed documents are vanity or cosmetic home pages, or valuable information created for a narrow, precisely defined audience, will not alter the fact that they appear as cyber-litter to the uninterested. One academic author recently referred to the Web as "multimedia mediocrity" because of the extreme variegation and lack of quality control over resources.

Search engines are inherently ill-equipped to deal with this problem. They work most efficiently when information is indexed, graded and categorized as they are posted. Since most searchable documents are now on the Web, and the Web didn't grow out of this philosophy, there is a definite practical limit to the performance one may expect of future search engines no matter how finely tuned.

This problem is worse than it needs to be because of inattention by the Web development community. For example, the 1993 robot exclusion standard could have provided a standard for "self-censorship" of Web sites. This would have gone a long way in making it easy for designers to exclude their less interesting, semi-private and derivative pages from indexing. As things now stand, the exclusion standard calls for a single file to be located in a server's main directory to handle all cyberspheres on a directory-by-directory basis. Even if this file were dynamically created from sublists of individual site owners, it is still too coarse to be very effective because site owners organize their directories for their own convenience, not for visiting robots. Not everything in a directory has the same value—even to the author and even if semantically related.

Similarly, exclusion of files could have been included in the earliest HTML standards by adding such meta-tags as "no index" and "commercial advertising" to the document specifications. Although this wouldn't disburden the network from all of the unnecessary packet traffic as millions of URL's were repeatedly visited without effect, it would at least remove a lot of the clutter for the search engines.

Search engines work with these constraints. While some additional effectiveness may be expected in such areas as indexing behavior (e.g., more sophisticated parsing and integration of <meta> and <title> tags with the index of the document body), it is unlikely that even the best-groomed search engine can satisfy all of our long-term needs. It appears as if the Web and the Internet continue to be over-indexed for the foreseeable future.

A partial solution is to develop personal software agents, information customization tools, and to introduce resource brand names to the Internet. All three are underway and will become important network resources in the next century.

Information Agency

The notion that a computer program might act faithfully on the owner's behalf is not new. Computer scientists such as Nicholas Negroponte and Alan Kay have toyed with such ideas for many years. However, it has only been in the past few years that such software agents have actually been deployed on the Internet.

Personal information agents act on behalf of their owner and take actions as the situation warrants. These agents (called software robots, softbots or, simply, bots) are computer programs that roam the Internet on behalf of their owners, ferreting information. If such agents are competent and trustworthy, they may free the owner from some of the labor involved in navigating/browsing the Internet and interacting with intermediate document-handling software such as search engines. If the agents are competent and trustworthy, that is. Therein lies the present

research rub and challenge.

Several general strategies for deploying software agents have evolved in recent years. Ironically, they are technically rooted in computer viruses—computer programs that either insert themselves into, attach to, or replace authorized programs prior to execution, and then replicate and perform some pernicious task. As with information agents, viruses also roam the networks on behalf of owners—though toward malevolent ends.

The following properties have been associated with software agents in the literature:

- Mobile—either migratory or nomadic
- Autonomous
- Self-initiating or proactive
- Social (communicate with both owners and other agents)
- Reactive to changes in external circumstances
- Persistent
- Capable of planning
- Capacity for beliefs, desires, intentions
- Rational
- Veracity
- Subservient

and, of course,

- Competent
- Trustworthy
- Benevolent

Agents of such stripes are now operational for both direct and indirect location, manipulation, monitoring and transmission of a variety of Internet resources within several Internet protocols. Such agents will be commonplace in the next decade.

Information Customization

Information customization, in the sense that we use it, is a useful complement to information agency. It has five basic characteristics: 1) it is always performed on the client side; 2) it is specifically designed to maximize information uptake, rather than filter or retrieve; 3) it "personalizes" documents by such techniques as extraction; 4) it is never done autonomously; and 5) the capabability of nonprescriptive, nonlinear document traversal is always added by the software. Condition (2) sets information customization apart from traditional information filtering and retrieval, while (4) sets it apart from client-side agency, and (5) would distinguish it from traditional nonlinear document traversal systems (e.g., hypertext).

In general, information customization software has the following sorts of architectural properties:

- Nonlinear traversal
- Nonprescriptive nonlinearity
- Interface based upon desktop (vs. application-specific) metaphor
- Noninsular—completely integrated with productivity applications
- Operates interactively in real time on documents
- Platform and document format independent

In short, information customization involves an interactive process whereby users would interactively and in real time control the means by which documents are reduced in size, or transformed into a more useful form, and dis-

played. Figure 2 illustrates this process in a proof-of-concept prototype called Cyberbrowser that behaves as either a stand-alone application or a browser-launchable spawnable peruser for either text or HTML documents.

Internet Brands and Brand Loyalty

The third prong of the next decade's attack on information overload is brand identification. Neither information agency nor information customization (not to mention search engines) will be able to handle the tidal wave of networked information. There must also be an entire network of information providers who will grade, rank, review, append, annotate, transfix, collect, and repackage Internet resources. It is with these providers that we will come to develop brands and brand loyalty for reliable, useful, current information and services.

In the case of the Web, information providers will come to be known for the quality and utility of the sites they manage. These will carry with them the digital imprimaturs of their hosts and patrons. These imprimaturs will bestow value on the documents of their sites in virtue of the widely acknowledged standards of their hosts and patrons, including the robustness of their review and screening, and their overall reputation as organizations. One may come to rely on the overall quality and reliability of information on the servers of corporation *X* and *Y*, but hold those of corporation *Z* in repute. Of course, the publishing and entertainment industries have worked this way for most of their histories. So far, the Web has yet to be so influenced. At this point, the best we can hope for is the carnival atmosphere of the many "top *n*%" sites.

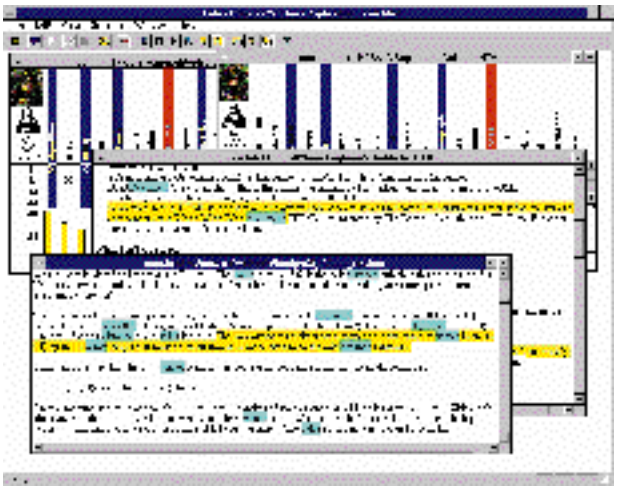

Figure 2. Our vision of the interface of an information customization prototype. In this case, automatic highlighting is employed so the user may view those portions of the documents that have been found relevant to current interests in context. Keywords (in blue) are identified and linked nonprescriptively by the client software independently of the HTML links.

Eventually, brand identification will also be incorporated into search engines. Over time, we'll come to rely more and more on network registries that will screen, review, rank, grade, or perhaps even ignore, URLs submitted by authors, publishers, distributors, and so forth. These registries will provide this service, in many cases for profit, for their customers and subscribers to add value to the documents. With secure http standards and the technology of network micro-transactions now in place, the future for this sort of digital commerce should be secure.

The Coming of Push-Phase Information Access

As important as information agency, information customization, and brand recognition will become, they alone will not be enough to harness the information access potential of the Internet. They are all predicated upon an information-pull strategy where the users, perhaps through autonomous software agents, seek to draw information to the user. We are now entering the push phase of network information access.

Currently, the paradigm is "solicited push" as individuals connect to various netcasting network information providers. With Pointcast (www.pointcast.com), users connect to a central digital transmitter that connects the end user to a variety of different information feeds (Reuters, Business Wire, *People* magazine, and so forth) integrated on Pointcast's server. On the client side, the Pointcast peruser operates as an autonomous window providing downloaded information from selected feeds. Pointcast follows in the tradition of the cable television industry by consolidating and distributing information (including advertising) from one distribution source to many subscribers.

Marimba Corporation's (www.marimba.com) approach to

solicited push is quite different. Here, an analogy with direct broadcast television satellite systems is closer. Marimba's proprietary client-server software, Castanet, allows the end user to connect to an arbitrary number of third-party server transmitters from the client side. The connection between a Castanet client "tuner" and each of the Castanet server transmitters is called a "channel." In basic terms the channel permits network access to some server's file structure. Where Pointcast is a one-to-many network transmitter, Castanet is many-to-many. (see Figure 3).

Castanet and Pointcast are also interesting extensions of the traditional network desktop metaphor that has to this point been browser-centric. On this account, the network point-of-contact is a single network client (e.g., Netscape for the Web) from which all other applications are spawned (in Netscape's case either through its internal launchpad or as a plug-in). While both Pointcast and Castanet are browser independent, Castanet is the most supervenient with respect to the browser—each channel-application 1) performs autonomous local processing; 2) enjoys persistent, client-resident storage; and 3) updates both data and program upgrades differentially, by sending only new or changed files. Further, with Castanet, differential updating is client-initiated, setting it apart from Pointcast and facilitating authorized penetration of server firewalls. Castanet is a very real, robust departure from the browser-centric approach to client-connectivity—exactly what is needed if it is to succeed as a rival technology to Web browsers. This point was apparently poorly understood by the developers of Hot Java. As an aside, plans are already underway for Netscape, Marimba and Pointcast to collaborate on meta-push technology that will integrate all three technologies and services in a common intranet client, Constellation.

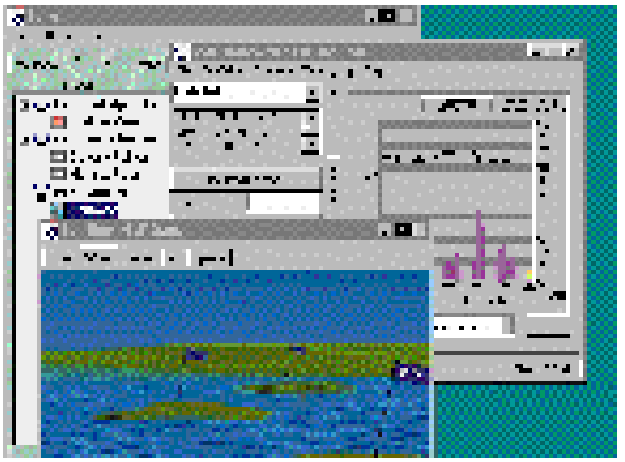

Figure 3. Marimba's Castanet Client. Note three channels are established, the first and third of which are active in separate windows. Interactivity between client and server is automatic and autonomous for each channel.

So we predict solicited push network environments are a done deal for the next decade's cyberspace—they will evolve in parallel with the resources of the World-Wide Web and its successors to fill a specific niche in tomorrow's information infrastructure. The open question is how long it will take before the unsolicited push networks evolve. That realization will produce a full-blown development effort for the third phase of network information access—the "repel phase." More on that in a future column.

For Further Reading

• **Information Agency**. An excellent starting point, even if a bit dated, is the July, 1994 special issue of *Communications*. Many of the key players in the field contributed to this issue.

Another useful overview is Fah-Chun Cheong's *Internet Agents: Spiders, Wanderers, Brokers, and Bots* (New Riders, 1996).

• **Information Customization**. Our early ideas were outlined in a two-part series in the September and October, 1994 issues of IEEE *Computer* entitled "Customizing Information," and an article entitled "The Challenge of Customizing Cybermedia," a draft of which is available online via my home page at www.acm.org/~hlb/.

•An insightful look into the history of computer viruses may be found in Peter Denning's *Computers Under Attack: Intruders, Worms and Viruses* (ACM Press, 1990). ◻

Hal Berghel (www.acm. org/~hlb/) is a professor of computer science at the University of Arkansas.