Infoglut

Overload of cheap information threatens our ability to function in networks; value-recognizing architectures promise significant help.

Nearly 25 years ago, I wrote a *Communications* "President's Letter" column entitled "Electronic Junk" [1]. At the time, the Internet (then not a widely used term) was only 200 nodes, but there were already signs in our local networks that information overload would be a chronic disease.

Of my own situation I wrote: "In one day I typically receive 5–10 pieces of regular junk mail, 15–25 regular letters, 5 pieces of campus mail, 5 reports or documents (not all technical), 5–10 incoming phone calls, 10–20 local electronic messages, and 10–20 external electronic messages. Although many of these messages are discarded or forwarded to others for handling, much of my time is required to skim and dispatch them. Although I save only those reports whose titles and abstracts sound very interesting, the pile of unread reports continues to grow on the table in my office." (How quaint the terminology: mail and electronic messages instead of postal mail and email.)

Then, looking to the future, I wrote with trepidation: "Beyond the riptide of normal business mail lies a tidal wave of electronic junk mail. It is now trivial for any user to send copies of virtually any document to large sets of others. ... The growth of new networks such as CSNET and USENET only adds to the heights of the waves of materials that try to flood any given person's mailbox. It is clear that some attention must be paid to the processes of receiving informa-

tion, and preventing unwanted reception."

In the past quarter-century, the tsunami arrived. Most of us routinely see hundreds of daily email messages. Many come with attached documents or long email threads that someone wants us to read and act on. The Web is so large that we can find things only with powerful search tools like Google, and even then we often find tens of thousands of matches to our keywords. Since we almost always select from the first 10 items on the match list, we're often left with a nagging suspicion that we missed something really useful farther down the list. Despite our large investments in spam filters, our mailboxes keep filling up with junk, drug offers, pornography, sales pitches, attempted virus implantations, scams, and phishing expeditions. I read an estimate recently that spam is now well over 50% of the email transmitted on the Internet and that it costs the U.S. econ-

omy over $10B annually in lost productivity.

I mentioned in my letter eight technological aids to limit a user's received information flow: hierarchical organization of mailboxes, separate private mailboxes, special forms of delivery, content filters, importance numbers, document skimmers, quality certification, and bid-ask threshold reception. All but the last are used today, but even so the tide of filtered information is overwhelming. The Internet easily defeats advanced filters, delivering millions of words per second to brains that can process only 10 words per second.

The Internet technology has given us a tragedy of the commons: anyone can direct large amounts of information to me at virtually no cost to himself. A single advertiser can send a message to a million people, but does not see that the aggregate cost of those recipients spending five seconds each to delete it adds up to 58 days of lost productivity. A single user can waste an hour a day simply deleting 700 unwanted email messages.

Data Smog

The Internet is not the only overwhelming source of data. In 1997, David Shenk published *Data Smog*, documenting the full extent of the information overload problem [7]. In addition to the Internet, we are offered information from television, radio, fax, phones (regular calls, telemarketing, text and instant messaging, pictures, videos), advertising, and personalized junk mail such as pre-approved credit card applications that must be shredded to avoid the possibility of identity theft. In many of these media, as in the Internet, we must cope with the scourges of spam, scams, viruses, hijacks, and phishing, all adding to the overload.

In 1970, psychologist Stanley Milgram, studying people's reactions to the overloads of city living, cited six coping strategies: spending less time on each input, disregarding inputs, shifting the burden to others, blocking reception, filtering, and creating specialized institutions to offload the work [6]. These strategies are uncannily similar to the ways we deal with Internet overload: we don't read carefully, we disregard, we hand off tasks to others, we block reception, we filter, and we create institutions to share the burden (for example, spam-blocking services). Milgram said, "City life as we experience it constitutes a set of encounters with overload, and of resultant adaptations." He could have said the same sentence with "digital media" replacing "City life."

Shenk exhibits a curve, measured by psychologists, showing that the observed information-processing rate of the brain first rises, then peaks and declines with increasing rates of requests for processing. When this happens in a computer system or network, we call it thrashing. Our brains thrash when overwhelmed with too many incoming bits.

The mismatch between our capacity to process information and the rate new information arrives takes a heavy toll. When we are persistently overwhelmed, many of us feel highly stressed and experience stress-related health problems. We worry that our children, mesmerized by television and video games, don't learn to think for themselves. We become detached and uninvolved. We lose our ability to focus deeply on one item—witness the increasing number of individuals afflicted by attention deficit disorder. According to polls, we are remarkably uninformed about current events even though surrounded with 24x7 news feeds.

Paradoxically, Shenk says, even when we see that technology is the source of these afflictions, we look to more technology for the cures. We want faster search engines, not a smaller Web. We want smarter spam filters, not economic disincentives to spam. We want to record every bit of information that we send or receive even though we doubt that anyone else cares. In the belief that the technology gives us a "voice," we have created 50 million blogs and 5 billion Web pages—and then we wonder if anyone really notices. Our love of technology and belief in its redemptive powers is as strong today as it was a century ago [8].

In short, the Internet is really a small part of the total picture of information glut and our coping strategies are much the same for all the forms of glut.

Valued Information at the Right Time

Are there ways we can organize our technology to help us out of the information glut? A new approach is gathering momentum. It's called "valued information at the right time," abbreviated VIRT. Rick Hayes-Roth has been one of its chief proponents [4]. It's not so much about technology as it is about deciding which information is of value and to whom, and then configuring the technology

information to flow from a supplier to a consumer:

• Push: The supplier initiates the action with an offer. Broadcasts, standard email, spam, and subscriber distributions fall into this category. The consumer may not always accept the offered transmission; for example, a spam blocker stops it.

• Pull: The consumer initiates the action with a request. Queries in a language such as SQL (for databases) and Google searches

push data once that consumers would otherwise have to pull with repeated queries.

Dieter Gawlick of Oracle gives this example. Suppose you know of an auto broker with a database of available cars listed by participating dealers. You can periodically log in to the broker's server and query for cars of interest. However, a car of interest can be offered and then sold between your logins. The service becomes more valuable to you if the auto

The push and smart push configurations differ significantly in their ability to ease infoglut.

accordingly. It's bringing a human dimension back to an inhumane consequence of information technology.

At its highest level, a distributed communication system is a network connecting a set of information publishers to a set of information consumers. The ideal network delivers a bit stream to each consumer comprising just the bits of most value to that consumer in addressing current concerns or interests. This is the core of the VIRT idea.

To meet this ideal, there must be a way for consumers to reveal what is most valuable to them and for the network to adjust the flow to each accordingly. This is done best with consumer-supplied "conditions of interest" and networks configured for "smart push," as discussed shortly.

Two kinds of action cause

are prime examples.

Pull strategies will be the better choice for a consumer satisfied by a single snapshot of the data. More often than not, however, consumers try their queries repeatedly, searching for a satisfactory snapshot. Such consumers will find it valuable to put their query on file and be notified automatically if it becomes true. In the parlance of the VIRT world, such a consumer creates a subscription and the publisher pushes information as often as it is available to fulfill that subscription. Subscriptions are entrusted to subscribers, special roles that initiate flows from publisher to consumers. Although publishers and consumers can be subscribers, the most common situation is the independent subscriber, which acts like a broker between publishers and consumers. Subscribers

broker acts like a subscriber relative to the dealers (who publish car offers). You can file a condition of interest (COI) with the broker, who will notify you immediately when a dealer posts a matching car. A condition of interest is a statement of the form, "I am in the market for a car in PRICE RANGE having one of the colors LIST and the options LIST." The DVD distributor Netflix operates in a similar way.

Conditions of interest express what consumers consider most valuable to them. A system configured to send to consumers only the data that satisfies their previously filed conditions of interest is called smart push. A well-designed smart-push system never sends worthless information.

Complex Event Processing is a technology for smart push [5]. The idea is to express a condition

of interest as a complex combination of observable events. Some events can be generated by "triggers," which are autonomous processes that continuously scan changing data sets for specific patterns.

The push and smart push configurations differ significantly in their ability to ease infoglut. In the simple push configuration (see Figure 1), the COI subscribers take the form of filters located with the consumers, discarding unwanted data. This configuration can fall victim to network congestion: the network carries copious data of value to no one. Many critical networks do not have extra bandwidth [2]. Some publishers attempt to mitigate network overload by dividing the data stream into many "channels" and asking subscribers to specify channels of interest.

In the smart push configuration (see Figure 2), the COI subscriber function is located at the data server, where it can scan deeply into the database to detect patterns meeting the COI and then push selected data back to the consumer. A COI generator agent is placed with the consumer; it monitors the consumer's context, actions, and words; and it generates COI expressions, which it sends to the COI detector.

Rick Hayes-Roth illustrates the dramatic difference between these configurations with an example of a helicopter pilot who plans a low-risk route through a war zone [4]. Before starting, the pilot cre-

ates a flight plan that avoids storm cells and air defense positions. The pilot will deviate only on learning of changes in storm and defense positions, as well as movements of other aircraft, that intersect the flight path. Various

Figure 1. Simple push configuration.

other technologies (weather observation, radar) track storm movements, anti-aircraft positions, and other aircraft through the entire region. Of all this information, however, only the data that would cause the pilot to deviate from planned flight plan will be valuable. Which configuration assures that only the most valuable data is actually sent?

Hayes-Roth considers a flight path through a region 200km on a side. Sensor resolution in the region is 1km, giving 40,000 grid points. Vertically, data is available at 500m intervals from altitude 0km to 6km, a total of 13 altitude coordinates. That gives 520K grid points in the 3D volume. Forecasts of 10 variables are tracked at each grid point, giving 5.2M data values in the volume; these forecasts are updated every 30 minutes. The flight is sched-

uled for 4.5 hours, giving 10 update times. Thus the total size of the data space is approximately 52M values.

In a push environment, the sensors and updaters send new information to the pilot when-
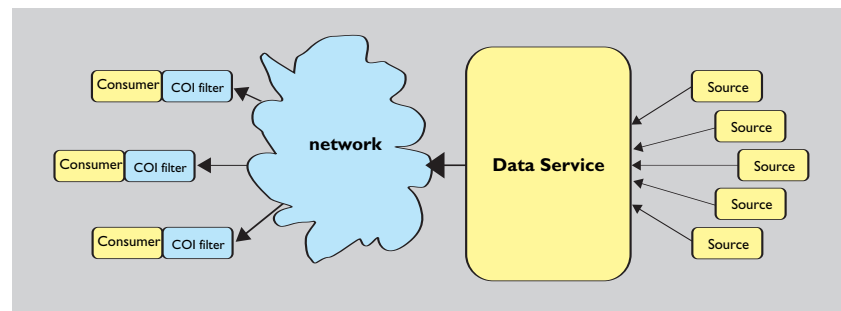
ever they get it; so during the 4.5-hour flight, the pilot would receive all 52M values. The pilot will not see all these values because he set his local COI filter to discard data more than 5km away from the flight path and data that changes less than 5% from previous reading. Even if the filters remove 99% of the offered values, the remaining 1% (520K potentially relevant values) exceed the pilot's capacity to make sense of them. Not only that, but the 99% of values discarded wasted bandwidth and prevented other pilots from getting valuable data on time.

In a smart push environment, the pilot tells his local COI generator agent that data outside some radius of the planned flight path is irrelevant and that alerts should be given only about variables that deviate enough from prior values to cause a change of flight plan. The local COI generator builds a

COI expression and sends it to the corresponding COI detector at the data server. The pilot knows from experience that he is not likely to see more than five alerts on the whole flight, well within his processing capacity. If each alert is

Figure 2. Smart push configuration.

accompanied by 100 data values (to update the display), the five expected alerts present about 100,000 times less data than in the simple push environment. These differences are significant and are very attractive to our pilot.

In this system, value is incorporated into the design through the COI generator agent attached to the consumer and the corresponding COI detector attached to the data server. The user's context and intentions define which values are relevant and should be pushed by the data server.

Workflow systems are another category of architectures that use the VIRT principle. These systems track commitments in a network of people who are engaged in a standardized work process. Information flows between people only as they make requests and fulfill promises. Since every

information flow is essential to the work process, all flows in a workflow system are of high value. Unfortunately, many real networks, such as hastily formed networks, do not have a well-defined set of workflows and can

benefit only marginally from a workflow approach.

Conclusion

The information glut problem we experience in our digital networks is part of a much larger information glut problem throughout all communication media. The problem has arisen because technologies are able to help us generate information much faster than our individual capacity to process it. Thus much information is lost or ignored, and as users we become overwhelmed, frustrated, and detached.

As we build more automated sensing and data collection environments, the overloads will only get worse. To stanch the flow, we must fill a gaping hole in networking technology—its architecture does not consider the value or relevance of information to a potential receiver. System architectures incorporating this princi-

ple—VIRT technologies—can limit information flows to individual users without losing effectiveness. Even in data-dense environments, a smart push VIRT strategy can reduce the flow by five or more orders of magnitude, enough to match the user's processing rate and achieve a significant advantage in resource usage and productivity. Intel chairman Andy Grove told us that any technology with a 10x (or more) advantage over the current is potentially disruptive [3]. Keep your eye on VIRT. **C**
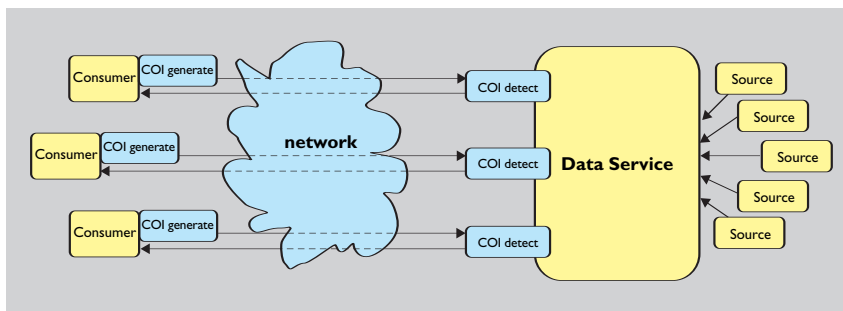
References
1. Denning, P.J. Electronic junk. *Commun. ACM 25*, 3 (Mar. 1982), 163–165.
2. Denning, P.J. Hastily formed networks. *Commun. ACM 49*, 4 (Apr. 2006), 15–20.
3. Grove, A. Only the paranoid survive. *Currency* (1996).
4. Hayes-Roth, F. Two theories of process design for information superiority: Smart pull vs. smart push. In *Proceedings of the Command and Control Research and Technology Symposium: The State of the Art and the State of the Practice*. San Diego, CA, U.S. Department of Defense, Command and Control Research Program (CCRP), 2006; www.nps.edu/cebrowski/Docs/06reports/CI-06-001.pdf.
5. Luckham, D. *The Power of Events: An Introduction to Complex Event Processing in Distributed Enterprise Systems*. Addison-Wesley Professional, 2002.
6. Milgram, S. The experience of living in cities. *Science* (Mar. 3, 1970), 1461–1468.
7. Shenk, D. *Data Smog*. Harper Collins, 1997/1998.
8. Walter, D. *Today Then: America's Best Minds Look 100 Years into the Future on the Occasion of the 1893 World's Columbian Exposition*. American World Geographic Publishing, 1992.

Peter J. Denning (pjd@nps.edu) is the director of the Cebrowski Institute for information and innovation and superiority at the Naval Postgraduate School in Monterey, CA, and is a past president of ACM.