We Digital Sensemakers

MARK STEFIK

> The creative challenge is not so much in gathering information as it is in asking the right questions.
> Joshua Lederberg

Sensemaking is the process by which we go about understanding the world. It is as natural as breathing and eating. Everyone does it. Sensemaking employs a raft of cognitive activities, including perceiving and interpreting sensory data, formulating and using information, and managing attention. It also employs social activities such as sharing, recommending, critiquing, and discussing. The promise of digital and social sensemaking is to radically improve our ability to make sense of information.

"Digital sensemaking" is sensemaking mediated by a digital information infrastructure, such as today's web and search engines. While the amount of information continues to expand rapidly, our innate human capabilities to make use of it are approximately fixed. Digital sensemaking counters the growth of information by harnessing ever faster computing. Web search engines have greatly improved our ability to find information. However, tools for sensemaking still fall far short of their potential. Sensemaking on the web is often frustrating and onerous, requiring one to wade through off-topic and poorly written pages of questionable authority.

Even professional sensemakers experience failure and frustration with current tools. Intelligence analysts are the jet pilots of sensemaking, addressing the most extreme professional challenges. Their work involves requesting and otherwise collecting an immense amount of information, sorting through it to identify relevant pieces, and constructing and maintaining an

understanding of international dynamics for tactical and strategic purposes. Yet they failed, for example, to anticipate the Yom Kippur War in 1973 or the collapse of the Soviet Union in 1991, and in the months before the 2003 invasion of Iraq reached the apparently false conclusion that that country possessed weapons of mass destruction. Despite much expense and many organizational reforms, the information gathered by the intelligence community extends far beyond its ability to make use of it.

Two themes guide our ongoing pursuit of quality and ease in sensemaking. The first is the challenge of finding not just the right answers but the right questions. The second is the recognition that there is more power in sensemaking when it is cast as a social activity than when it is seen as an individual pursuit.

Although digital sensemaking today is mostly a solitary activity, social-media approaches are now emerging that may radically change the experience of digital sensemaking. Social sensemaking will counter the proliferation of information sources of varying quality with the collective knowledge and judgment of people, helping us to combine our efforts to evaluate the quality and relevance of information, to develop shared understanding, and to put information to use.

1. Information and Attention

"A wealth of information creates a poverty of attention," Herb Simon once observed (1971, 40). Information and attention are the key resources that we manage in sensemaking. Publishers and professional sensemakers are acutely aware that far more information is available than any of us can consume. Our *collective* consumption of information can be described by a long-tail distribution.[1] How we consume information *individually*, however, is better understood in terms of "information diets." This term refers to the information that we consume across different categories, including topics in the news, professional interests, hobbies, and entertainment media. The categories are different for each person. An information diet can be represented as a list of subject areas with figures indicating how much of our time or attention is allocated to each. The total allocations add up to 100 percent of our available time.

Although popularity curves summarize our information consumption in the aggregate, they do not describe us as individuals. For example, the day's most widely consumed news stories may constitute only a minor share of my daily information diet. Except for some teenagers, there are relatively few

people who follow the dictates of popular taste so rigorously that their personal top story, favorite piece of music, and so on correspond to the collective favorites.

Information diets are different for each person. What we have in common is the frustration of clumsy sensemaking services. Current search tools and news services are optimized to serve the head of the long tail. Unfortunately, this information infrastructure, optimized to serve us in the aggregate, does not serve us very well as individuals.

2. Three Challenges for Digital Sensemaking

For our ongoing information needs, the challenge is to track new information that is relevant and important to us. New information becomes available from many different sources. Web search engines are not ideal for satisfying an information diet. They do not enable us easily to focus on a subject area or topic, and typing "What's new?" into a web search box does not yield a useful response. Web search engines generally make little note about whether content is fresh or stale. They favor old information. They prioritize search results using interpage linking structures to estimate authoritativeness and aggregate popularity. Consequently, a web page usually will not be ranked high enough to come into popular view until enough links are made to it, which is probably long after it was new. In contrast, mainstream news services focus on fresh and popular information. The information is organized into broad categories such as "business," "national," "international," "entertainment," and "sports." Such broad categories do not serve the specialized interests in our information diets.

Given limited time and an ongoing concern that they will miss something important, many experienced information consumers employ two kinds of tools that cover topical information from much farther down the tail: RSS feed readers, which allow them to subscribe to professional news feeds and blogs, and news alert services, which filter articles from thousands of sources based on search terms. There are now hundreds of thousands of such sites on the web. Both approaches provide levers for managing attention, balancing information overload against the risk of missing important information, yet neither provides enough help in sorting through the tide with an eye to quality and authority. Users of feed readers can control which sources they pay attention to. They scan titles of new articles on a regular basis, but articles on narrow topics still represent a small fraction of the information on broad feeds. Since even two or three feeds deliver more information than most peo-

ple have time to scan, they may miss articles that appear only in other feeds. Alert services have different leverage and different problems. If people use search terms that match a broad range of topics, they again face an overload of incidental matches and stories from dubious sources. If they narrow their terms, they risk missing important, related information. As with RSS feeds, people generally do not subscribe to more than two or three alerts.

This brings us to our first challenge for sensemaking and information foraging: developing better approaches for tracking new information on the core interests of our information diets.

Information just beyond the edges of our interests constitutes our "information frontiers." We may know people who are familiar with it, but it is over the horizon and beyond the reach of our personal radar. The frontiers in professional fields are topics from related and nearby fields. In community news they often include happenings in neighboring communities. Information frontiers in business and technology may reveal new developments that bring change and opportunity. Exploring frontier information helps in spotting new trends.

As a director at the Institute for the Future, Paul Saffo analyzes technology and business futures. In an interview about their forward-looking process he said:

> When you are mapping out technology horizons and making forecasts, you focus on opportunities at the intersections of fields. If you want to innovate, look for the edges. The fastest way to find an innovation is to make a connection across disciplines that everybody else has missed. (Stefik and Stefik 2004, 167–68)

Saffo's interest in the frontiers or edges of a field brings to mind Ronald Burt's ideas about structural holes in social structures (Burt 2004). People attend mainly to ideas circulating within their group. This leaves "holes" in the flow of information between groups. Burt's hypothesis is that new ideas emerge from *synthesis across groups*. People who are connected across groups become familiar with multiple ways of thinking and thus are better positioned to detect opportunities and synthesize ideas. In short, they have an advantage of vision and use it to broker ideas. Frontiers are challenging because the amount of information on our frontiers is larger than the body of information in our main focus and it is less familiar to us.[2] Consequently, we need more help in allocating some of our scarce attention to scan our information frontiers.

Our second challenge, then, is finding better approaches for gleaning information from beyond our information frontiers. We occasionally need to learn about topics that have not previously been of interest. We may, for example, be considering the purchase of a new kind of appliance. Or we may need to substitute for a coworker on leave whose specialty differs from our own. When a family member develops a health problem, learning about treatments and services may become a sudden, urgent priority.

This brings us to a third challenge for information foragers: coming up with better approaches to support *understanding* in an unfamiliar subject area.

In summary, the three challenges for digital sensemaking are information tracking (keeping up with core interests), information discovery (discovering information from our information frontiers), and information understanding (making sense of subject areas that are new to us). The rest of this chapter takes each of these challenges in turn, considering the nature of each challenge and the emerging technologies that can radically improve our experiences as digital sensemakers.

3. Tracking Information in Our Core Interests

In a typical information-tracking scenario, sensemakers have access to materials with information on their core topics. New materials, arriving from multiple sources, are not categorized by subtopic and may include information beyond our information diets. Levels of authoritativeness may vary. The challenge is to classify the new materials at fine grain by subtopic and to quantify one's degree of interest in order to prioritize articles and allocate attention.

An automatic approach for improving information tracking must address three key subproblems: developing a useful topical structure, organizing new information by topic, and presenting articles within each topic in an appropriate order. The following discussion is based on our ongoing experience with three generations of social-indexing systems that we have built.

3.1. *Topics in Books*

We begin with the familiar example of books—which often include tables of contents and back-of-the-book indexes. A table of contents affords an overview of the information presented in an order useful for reading. An index allows for piecemeal access to information, according to our immediate needs, based on an expert's articulation of the book's important topics. Both embody judgments about how people will use the information in the book.

TABLE 1. DENSITY OF INDEX ENTRIES IN SELECTED BOOKS

Book	# pages indexed	# index entries	Index entries per page	Words per page	Words in book	Words per index entry
Open Innovation (Chesbrough)	195	448	2.29	390	76,050	169
Crossing the Chasm (Moore)	215	560	2.60	429	92,235	165
Problem-Solving Methods in Artificial Intelligence (Nilsson)	240	640	2.67	429	102,960	160
The Tipping Point (Gladwell)	280	630	2.25	310	86,800	137
Biohazard (Alibek)	292	832	2.84	407	118,844	143
The Psychology of Human-Computer Interaction (Card, Moran, and Newell)	431	620	1.43	350	150,850	243
The World is Flat (Friedman)	469	1,400	2.98	420	196,980	140
The Dream Machine (Waldrop)	472	1,440	3.05	559	263,848	183
Peasants into Frenchmen (Weber)	569	1,870	3.28	516	293,604	157
Applied Cryptography (Schneier)	620	2,140	3.45	580	359,600	160
R&D for Industry (Graham and Pruitt)	623	1,890	3.03	369	229,887	122
Readings in Information Visualization (Card, Mackinlay, and Shneiderman)	640	2,448	3.8	700	448,000	183
Introduction to Knowledge Systems (Stefik)	775	1,544	1.99	500	387,500	250
The Notebooks of Leonardo Da Vinci (MacCurdy)	1,186	2,970	2.5	400	474,400	159

Table 1 presents data about index entries from a set of books selected from my work office one afternoon. Some were academic and discursive, some were technical, and others were popular business books. The counts of words and index entries per page were determined by averaging over several sampled pages. The number of words per page (from 400 to 700) varied according to several factors, including the size of the page and of the type and the abundance of figures, tables, code, and headings. Index entries were counted at all levels. The number of entries ranged from about 2 to 3.5 per page or, taking into account variations in the number of words per page, one for roughly every 166 words. Although some index entries cite only a single page, most refer the reader to several pages, the average being about four. These data sug-

gest that indexers tend to tag content for indexing about every 40 words. This is about one or two tags per short paragraph. The data also show that each page in a book is cited in connection with eight to twelve topics. The index thus provides a relatively fine-grain tool for searching a book by "topic."[3]

3.2. *Problems with Automatic Indexes*

Various approaches to automatic indexing have previously received research attention, especially indexes based on concordances. A concordance is an alphabetized list of the words and phrases in a document together with their immediate contexts. Concordances can be compiled automatically, sometimes using linguistic techniques for phrase selection and normalization.

For purposes of information tracking, however, concordances fall short because their articulation of subtopics is not informed by domain expertise or historical experience. Unable to distinguish between the important and the trivial, they fail to identify and carve material along useful ontological and topical "joints."

3.3. *Generating Topic Models*

Although a book index is a good starting point, one inherent limitation is that it is static. It is prepared when a book is created and is frozen in time. This is fine for books but insufficient for dynamic information from online sources. What is needed is an automatic approach to extend topical indexing to new material.

We have developed an approach to this problem called *index extrapolation*. Index extrapolation starts with example pages for each topic, provided by human curators. The topics and their example pages are used as training information to bootstrap an evergreen index. Our machine-learning approach develops topic models and extends the index as new material is collected, as explained briefly in the sections that follow.

3.3.1. FINE-GRAINED TOPIC MODELS. Our approach to index extrapolation uses a hierarchical generate-and-test algorithm (Stefik 1995, 173). For each fine-grained topic, the index-extrapolation program analyzes the corresponding training pages and selects a set of "seed" words whose frequencies in the example pages are substantially higher than in a baseline set of pages sampled from many sources. Other words may be included as seeds when they are part of the topic's label or occur near a label word in the cited text.

The program then begins a systematic, combinatorial process to gener-

ate optimal queries, similar to the queries people use with web-based search engines. In index extrapolation, the optimal query candidates are expressions in a finite-state pattern language. The queries express subtopic recognition constraints in terms of four kinds of predicates: conjunctions, disjunctions, sequences, and ngrams (sequences of consecutive words). For example, a query might require that a particular seed word appear together with a particular three-word ngram or two words in a nonconsecutive sequence. Tens or hundreds of thousands of candidate queries are generated and matched against the training examples.[4] Candidate queries are rated according to whether they match the "on-topic" positive training examples and miss the "off-topic" negative training examples. A candidate query performs perfectly when it matches all of the positive examples and none of the negative examples. To choose a top query candidate when multiple candidates exhibit perfect performance, the evaluator also considers structural complexity and term overlap with the index label.[5]

The result of the machine-learning phase is an optimal query generated for every subtopic in the index. For example, in an early test of the approach using a book by a defector from the Soviet intelligence community, the book's index entry for the subtopic "Black Death" cited three pages among the several hundred pages in the book. Eighteen seed words were automatically selected, including "plague," "pesti," "yersinia," and "pandemic." About a thousand candidate queries were automatically generated and reported using the seed words. One candidate query required that a page include the word "plague," any word identified in a library as meaning "warfare," and either the word "bubonic" or the ngram "black death." Another required that a page include the word "plague," either the word "pandemic" or "rare," and either the word "yersinia" or "bubonic." The top-rated query, which required that a page contain either the word "bubonic" or the ngram "black death," was a perfect predictor on the training set without any false positives or false negatives, had some word overlap with the subtopic's index label, and had low structural complexity. Running over the entire book, the machine-learning program generated sharp patterns for each of the thousand or so subtopics in the index.

A more familiar example is the topic "housing crisis," which figured in an index about the "US Presidential Election 2008." Depending on the training examples, the optimal query computed by the system includes references to mortgages, housing, foreclosures, and bad loans. Our current social-indexing prototype has over two hundred indexes with several thousand topics.

3.3.2. COARSE-GRAINED TOPIC MODELS. Optimal queries are capable of identifying patterns of words that occur in the short paragraphs that cover the fine-grained topics we identified in book indexes. But books present information without much distraction. Web pages, by contrast, also contain words from advertisements, related articles, reader comments, and publisher notices. From the perspective of topic analysis, such additional material amounts to "noise" added to the information signal. The optimal query for finding fine-grained information across many web pages is vulnerable to being misled by this noise.

To cope with noisy information, we have found it useful to incorporate a

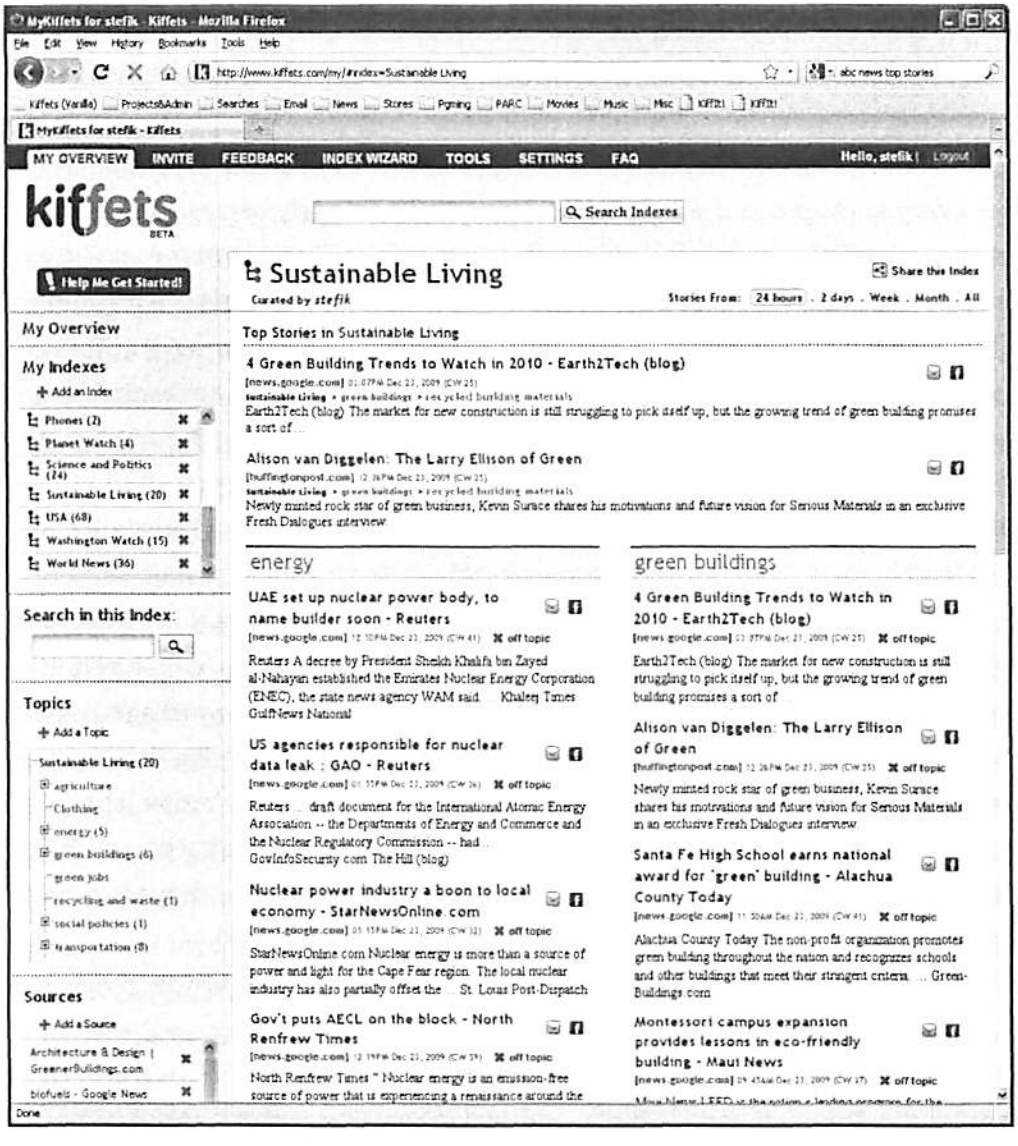

FIGURE 1. A social index about sustainable living. A tree of topics is shown on the left.

second, coarse-grained topic model that is less focused on small paragraphs. Again using the positive training examples, we compute a profile of the frequencies of the characteristic words for each topic. This word-population profile characterizes the kinds of words that are typically used in an article on a topic. Methods from information retrieval, such as cosine comparisons, can be used to compute a "distance" between two information sources based on their word usage.

These two topic models have opposite characteristics that make them powerful in combination. The optimal queries are capable of identifying fine-grained topics but are vulnerable to noise on a page. The word-population models are less precise with regard to topic identification, but are much less sensitive to noise. A web page about a sports story that has advertisements or a few story links related to the housing crisis will be reliably rejected by the word-population model for housing crisis.

Figure 1 shows an example of a social index about sustainable living. The topics used to organize the subject matter are shown in an alphabetical tree of topics on the left. The index curator has given a few training examples for each of the topics in the tree, and the system computed an optimal query for each topic. The system automatically organizes collected articles by topic.

3.4. *Keeping an Index Evergreen*

The index-extrapolation approach keeps an index open to new, arriving information. New pages are classified by subtopic by matching them against the queries. When a new page matches a query, it is registered as containing information on the corresponding subtopic. This approach is similar to information retrieval systems that use standing queries to retrieve new information. Index extrapolation differs from standing query systems in that the queries are generated automatically by machine learning rather than manually and that the topics are organized in a hierarchical topical index.

As a corpus grows, new pages may show up that should be included under a topic but are not matched by the query. When such pages are identified by a human curator or a voting process, they are logged as new, positive training examples.[6] When other new pages that are matched to a subtopic are judged as inappropriate for it, they are logged as new, negative training examples. Given such updates to the training sets, the machine-learning algorithm can be run again to revise the patterns. This tuning automatically improves the quality of the index going forward.

3.5. *Determining Degree of Interest*

Index-extrapolation technology addresses the first subproblem of the information discovery challenge: maintaining an evergreen index. We now turn to the second subproblem: determining a degree of interest for each information item. The degree of interest is used to rate and rank the articles or pages on a given topic, and to govern the display of the index information in a user interface. Compared to traditional media, social media offer fresh approaches to addressing the rating problem. Social media are distinguished from traditional media in their emphasis on social networks and their use of human feedback as a source of processing power.

3.5.1. RATING INFORMATION SOCIALLY. Digg pioneered a social-media approach to rating and ranking news stories based on the idea that people are the best judges of what news is important. Digg enables people to submit stories from the web or from news services and to vote on them. It also engages a social network of its readers. Members can subscribe to the stories that a friend or thought leader "diggs." The system maintains a list of current stories prioritized by their votes. As a story gets positive votes it rises on the list. If it gets negative votes, it drops down the list. To make the list responsive to recency, votes and article placement are adjusted for age so that older stories automatically drop and disappear. This approach to ranking stories initiates a positive feedback loop. As a story gets more votes, it rises in the list. As it rises in the list, it is more easily noticed. As it is more easily noticed, it can more easily attract votes. If a story gets onto the Digg front page, there is often a spike in the number of people noticing it. If a thought leader diggs a story, followers of the thought leader may also digg it, causing its rating to shoot upward. This kind of unregulated positive feedback has the potential for misuse and manipulation.

A warning about the workings of Digg's simple democratic voting system was sounded in 2006 when blogger Niall Kennedy noticed that many of the articles on Digg's front page were submitted by the same small group of Digg users voting for each other's stories. His analysis triggered a flurry of articles in various technology-oriented publications about the reliability of voting in social media. In 2007 there were multiple reports that cliques among Digg users were gaming the system in order to get articles on to the front page. A Cnet report, "The Big Digg Rig" by Elinor Mills, posted on December 4, 2006, described how some marketers were planting stories and paying people to promote them on Digg and other social-media sites. In response to this report,

Digg has modified the algorithms it uses to report, weigh, and count votes. Before considering methods for coping with voting problems, it is useful to look at some other issues relating to information tracking. Some typical criticisms of Digg are that it is too focused on technology topics and that articles on different topics are incoherently mixed together. There is an inherent challenge in satisfying multiple perspectives when a story is controversial or polarizing. If diverse communities used Digg, there could be a sustained tug-of-war over a controversial article; votes against would cancel the votes for, and the article would not rise in the popularity ranking.[7]

What kinds of articles appear on Digg? The category mix is indeed weighted toward technology. Even as the 2008 US presidential election was approaching, there were no Digg categories for politics or religion. At the time this was written, Digg had forty-nine classifications for articles, sorted under several general categories: Technology, Science, World & Business, Entertainment, Gaming, and Videos. Articles have just one classification, and it is established manually by the person submitting it. On the day I wrote this, the Digg front page had fifteen articles. Eight were about the technology industry, including one about Digg and several about the web. Three were about games. Two were about humorous online videos. Motor sports and international news had one article each. The list of top articles over the previous thirty days was a similar mixture, with mostly technology articles, including two about the iPhone. There was one article about a strange police arrest, and the rest were about videos. Religion and politics were not represented. Certain topics from down the tail are heavily covered (the network, operating systems, video games), presumably because they are important to the Digg community. Even in a specialized topic area such as World & Business, the articles are far from the mainstream, heavy on sensational stories and technology. This coverage suggests that the Digg community consists mainly of people under about twenty-two years of age who are deeply interested in computers, videos, and games. The particular topical focus of the Digg community is not bad, per se. It represents the votes of a self-selected population with similar interests.

In summary, current systems for rating news socially suffer from several problems. The dominance of cliques in promoting articles is a case of the tyranny of the minority. The suppression of controversial topics by vote canceling is a variant of the tyranny of the majority. Neither form of tyranny in voting is optimal for supporting information discovery across a community of diverse interests and values. This suggests that there is a flaw in the design assumption that populations are best served by aggregating all votes into a

single pool. What seems to be needed is an approach where users with different views are organized into multiple interest groups, each having fairly homogenous interests and values.

3.5.2. AUGMENTED INFORMATION COMMUNITIES. Organizing users into communities would make it possible for small groups and communities to explore their topics of interest and thus address the tyranny of the majority issue.[8] Each community would have its own index, covering topics in its subject area. Within a subject area, communities could pursue their particular segments of the long tail, rating materials according to their own values. In a technical subject area, professional groups might focus on advanced materials and amateur groups on introductory ones.

Most users would belong to multiple communities, corresponding to the core topics in their personal information diets. For example, a user might belong to one or more communities concerned with professional topics, a sports community related to a local team, a news community reflecting his or her political interests, a hobby-related community, and so on. Different communities could cover similar topics. For example, there might be "red," "blue," and "green" political communities, offering news and perspectives with, respectively, Republican, Democratic, and environmental slants. The placement and space allocated to displaying articles can also be governed by the community's voting practices.

While it may be useful to divide a population into communities of interest, it is also worthwhile to provide transparency across communities interested in related topics. Communities isolated from other worldviews risk becoming self-absorbed. A community whose interests or ratings became narrow and self-serving would probably fail to attract new members or much external attention. By enabling members of one community to see the topics and discussions of other communities, a discovery system can have a broadening influence.

3.5.3. STARTING A COMMUNITY INDEX. Dividing a population into communities introduces several interrelated issues. How do users join communities? How do they gain influence in them? How can a vote-based ranking system support discovery with rapid response to new information without being subject to the tyranny of cliques? How do communities keep from becoming too self-focused and narrow?

An online community may begin when a founding individual decides

to pursue some interest by starting a private index.[9] Acting as curator, the founder defines an initial set of online sources, such as new feeds, websites, or an online corpus. The index is bootstrapped either by starting with an index from another community or by starting from scratch, specifying subtopics and example articles. The index-extrapolation system automatically creates queries for each subtopic and finds further articles on them. At some point, the founder publicizes the index and opens up participation to like-minded individuals. As a community grows, members may be admitted at different levels. For example, an initial set of experts could be identified, with these "expert members" defined as thought leaders in the community. Experts' votes would have more influence in ranking articles than those of regular community members, and they could take on larger roles in maintaining the structure of the index by occasionally creating and editing topics. New members could gain expert status on the basis of social actions—referral, voting, recommendations, and so on.

Another category of users might be "harbingers." A harbinger is a community member who tends to be early, accurate, and prolific in identifying articles that the community ultimately ranks highly. Whereas experts might be appointed or elected, harbingers could be discovered automatically by tracking their submissions and votes over time. As their standing as accurate predictors of a community's interests and values is qualified, their votes could be given more weight than those of regular members. (Nonmember visitors could also use the index and read the recommended information but would have no say in rating articles.) If harbingers or experts were to have a streak of voting that was out of alignment with the community, their influence could automatically be decreased.[10] Having expert or harbinger status in one community would not give one similar status in a separate community.

3.6. *The Few, the Many, and the Machines*

This approach to the information-tracking challenge relies on three sources of power. The first is the hard work of the few, the experts who use their knowledge to curate and maintain a topical index. The second is the light work of the many, the people who identify and vote on disputed citations, influencing the training sets for tuning the patterns. The third is the tireless work of the machines—the index-extrapolation algorithms that automatically match the optimal queries against new pages to keep the index evergreen, the data-aggregation algorithms that combine the votes of the many to update the training sets, and the machine-learning algorithms that systematically cre-

ate topic models. Tireless by nature, computers can be massively deployed to meet the scale of the information and usage. These three sources of power are synergistic and fundamental to the design of social media.

4. Discovery on Our Information Frontiers

Discovery refers to finding materials on one's information frontiers, that is, in nearby subject areas. It is tempting to ignore the frontier. There is, as I have mentioned, more information there than in one's central field, and it is typically less important than that pertaining to core topics. Furthermore, the level of expertise of a sensemaker is lower at the frontier, with regard both to identifying good sources and to understanding topic structure. But there is a risk in not looking beyond one's core subject. Material that starts out on the frontier may become central as a field evolves, and early awareness of emerging trends can save the major expense of late remedies. Frontiers are resources for people interested in spotting trends arising at a field's edges.

As with information tracking, the value of discovery is better attention management. There are again three subproblems. The first is to identify frontier communities and their information. The second is to determine a degree of interest for ranking articles. The third is to relate frontier information to home topics.

4.1. *Identifying Information Frontiers*

In addressing information frontiers, we find it useful to focus on augmented communities as a level of structure and analysis for social networks. At the fine-grain level of individuals, a social network expresses relationships among people with common interests. At a coarser granularity, it expresses relationships among augmented communities that are interested in related subject areas.

Returning to Burt's analysis of communities and structural holes, each augmented community is intended to serve a fairly homogenous social group, in which members focus their attention on its core topics. Neighboring communities represent other fields or other groups. The technology for discovering information in a frontier is intended to provide a "vision advantage" that can be used for synthesizing new ideas and spotting trends.

When, in our model, the leaders of one community want to be made aware of relevant articles that another community finds interesting, they can designate it as a frontier neighbor. In a simple approach, candidates for neighbors might be found manually by searching a directory of communities. In

a more sophisticated approach, the multicommunity indexing system could suggest candidate neighbors using similarity measures that detect an overlap of interesting sources and articles between pairs of communities.

As a hypothetical example, a social index for topics related to "Music by Enya" might have as a neighbor a social index for topics related to "Music by Clannad," Clannad being a Celtic musical group that includes Enya's sister and other relatives. These indexes might connect to other social indexes on "Celtic Music" or "Irish Folk Music." For a geographic example,[11] suppose that there is a social index for the city of Palo Alto, California, where I work. Palo Alto's geographic neighbors include the cities of Mountain View, Los Altos, Menlo Park, and East Palo Alto, as well as Stanford University. For a medical example, a community interested in traditional Chinese medicine might focus on acupuncture and herbology. That community would be distinct from the myriad of "New Age" medical approaches in the West, although it might choose to designate such communities or one concerned with Ayurvedic (Indian) medicine as frontier neighbors. Networks of augmented communities could also be formed for sports, scientific studies, medicine and health subjects, religious subjects, and so on.

Reifying connections at the community grain creates a basis for tracking frontier topics and fostering cross-community information flows. Figure 2 portrays how an information community is located in a social network of other augmented communities, defining its information frontier. Overall, the social medium supports a galaxy of constellations of interlinked information communities.

Links to other perspectives can also be identified without requiring a curator to explicitly identify neighbors. By way of example, figure 3 shows a "front news page" that was computed on our prototype social-indexing system. The story about the swine flu was picked up in the "USA" index and organized under "Health and Safety/diseases/flu." Beneath the story are links to related topics that were identified automatically by the system. This calculation makes a second use of the word-population models discussed earlier. As the set of indexes and topics grows, the social-indexing system can compare the models for topics across all of the indexes. This makes it possible to identify cases where topics in different indexes are covering similar kinds of stories, albeit using different sources or with different user commentary. In this example, the system identified an index with a science perspective on the flu and also an index focused on China covering articles on the bird flu.

In summary, each augmented information community has its own index,

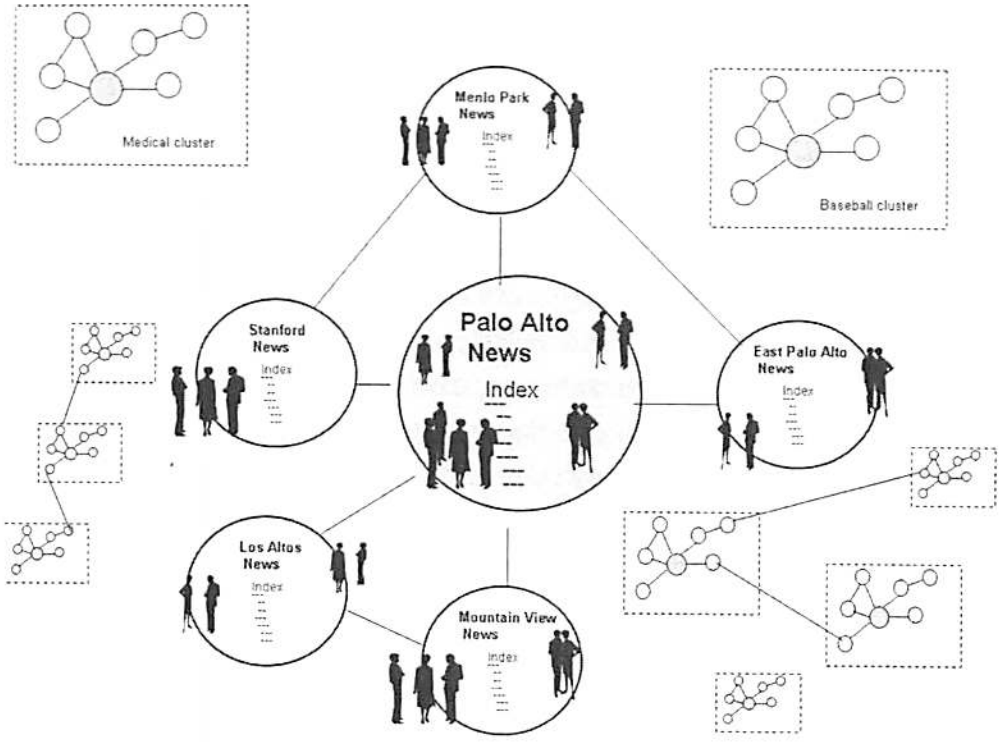

FIGURE 2. Each augmented community has its own information sources and social index and exists in a social network of augmented communities. The neighbors of a community provide a basis for computing its information frontier.

its own members, its own information sources, and its own ratings. Relationships between communities can be explicitly noted or automatically detected. The information resources from neighboring communities then become potential sources for discovering frontier information.

4.2. *Rating Frontier Information*

This brings us to the second subproblem for information prospecting. Given an information frontier, how can information be selected from the frontier and rated for interest? A useful starting point is the arrival of new information from the frontier communities. We propose using a degree-of-interest function that considers the level of interest an article generates in the frontier community, a distance metric quantifying the separation of the frontier community from the home community (such as the number of degrees of separation within the social network), and an indication of whether the article matches the topics in the home index.[12]

The neighboring community's sources and ranking systems thus provide for a first pass at identifying articles and a preliminary estimate of the degree-

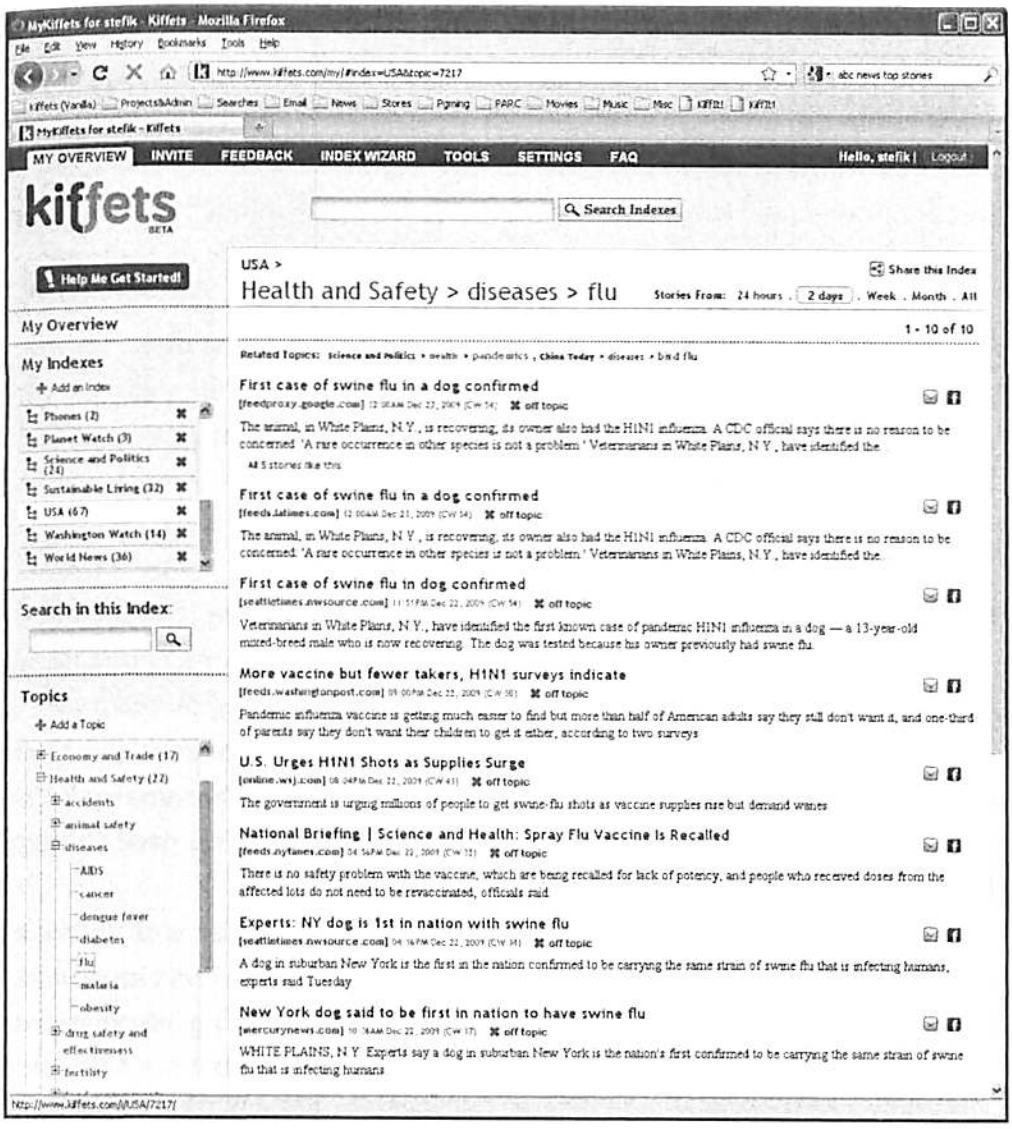

FIGURE 3. A news page in a social-indexing system. The top story is from the USA index, organized under "diseases" and "flu." Related topics below the story show perspectives on similar stories from related indexes

of-interest. Articles from neighboring communities can be assigned initial ratings (perhaps dependent on topics) reflecting the home community's ratings of earlier articles. These ratings can then be adjusted by voting and viewing response within the home community.

4.3. *Relating Frontier Information*

The third subproblem is to relate the frontier articles to home topics. Few articles from the frontier will be of universal interest in a home community. One approach is to automatically classify articles by the subtopics that they

match in the home index. In this way, articles can be routed to members of the home community according to their core topics of interest. In one approach, articles from the frontier get ranked and appear in topical indexes along with other articles from the home community's regular sources. As members of the community read articles on their core topics, highly rated frontier articles classified as being on the same topic compete for some of the display space.

In summary, the computational quality of social indexes provides new leverage for tracking frontier information for a community. The home community can rely on the expertise of its frontier communities to source and initially rate articles, and use its native index of topics to organize their presentation.

5. Supporting Understanding in New Subject Areas

Our third sensemaking and information-foraging challenge is understanding and orienting ourselves to information that is outside our usual personal information diets. Orientation refers to a process of getting familiar with a subject area, say, by learning about its topical structure, main results, and best references in order to answer questions important to the sensemaker.[13] The understanding and orientation challenge arises whenever we need to learn about something completely new.

This challenge relates to an old chestnut about struggles with information retrieval systems. How do we get the right answers if we don't know what questions to ask? How do we know what to ask for in retrieving information if we don't know what information is out there? How can we tell the difference between good and bad sources of information?

To explore the nature of this challenge, we consider again the fundamental properties of a good social index. An index provides a layered organization of topics. A good index embodies expert judgments about how people in the community understand the information. Index topics are somewhat like the "important questions" of a subject area. The structure of topics describes how people have found it useful to organize that area. The cited and ranked articles under each subtopic reflect a community's judgments about the best sources and approved answers for each subtopic. An index itself can be designed with some overview subtopics that serve specifically for orientation. Following this line of thought, *the challenge of orientation is largely addressed by providing a sensemaker with a good index.* Figure 1 shows a topical index related to sustainable living. Using it, a person new to the subject area can explore topics on sustainable agriculture, clothing, energy,

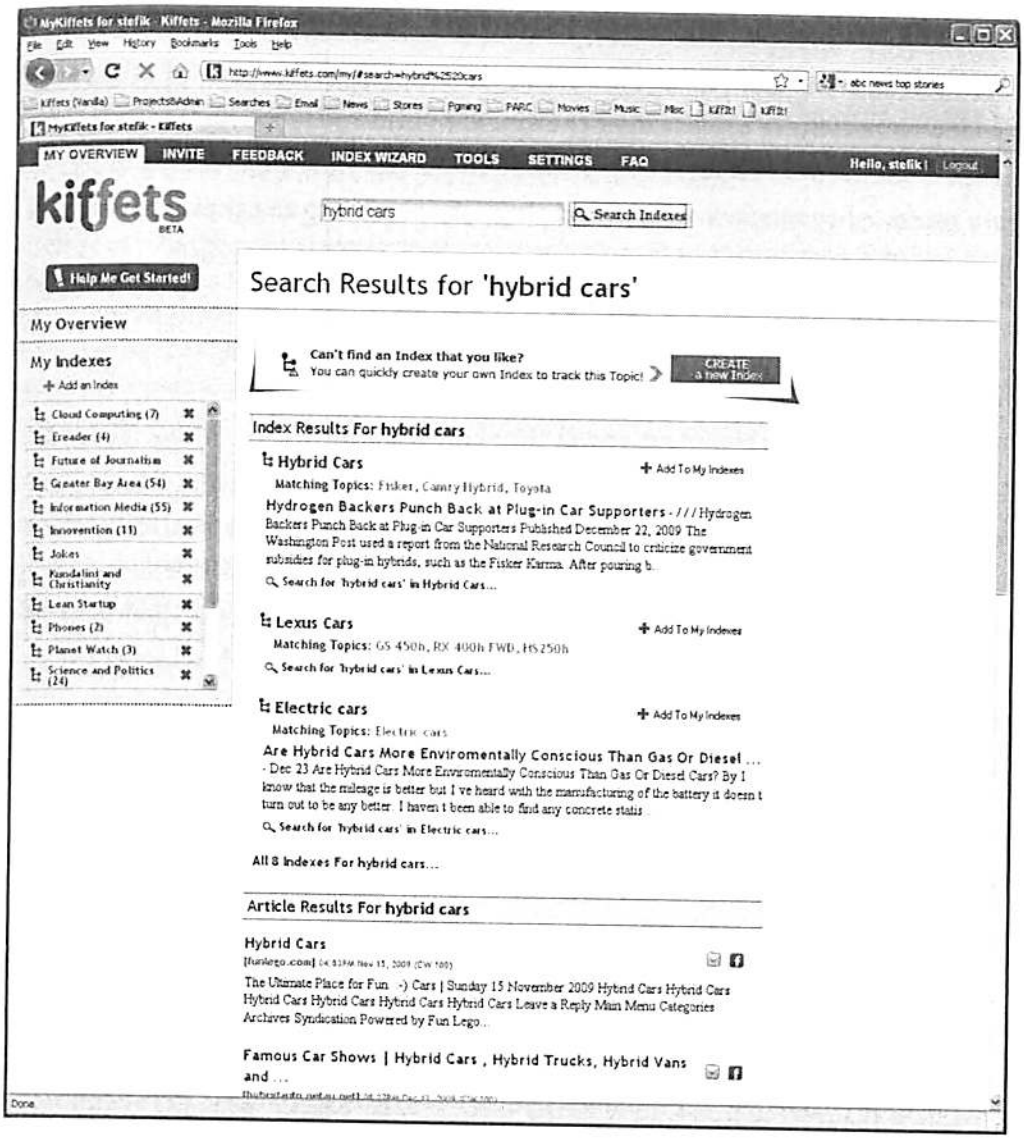

FIGURE 4. Finding relevant indexes. A search for "hybrid cars" returns a list of related indexes—ranging from indexes dedicated to the topic to indexes for particular manufacturers (with subtopics on their hybrid models) to indexes focusing on the technology of hybrid cars.

social policies, and so on. In short, the index itself is a guide to questions and answers about sustainable living.

Suppose that a person does not know which index to use to get started. Figure 4 shows how our current prototype helps one identify a suitable index. In this case, the person wants to find out about hybrid cars and enters a search query into the system. The system presents two sets of results—a list of indexes and their matching topics, and a selection of matching articles. The system shows indexes for hybrid cars, Lexus cars, electric cars, and others. For each index, it gives the best matching topics and a sample article. The top

index, "Hybrid Cars," is all about hybrid cars and has topics organized by manufacturer and model. The index "Lexus Cars" is about Lexus cars in general; its closest matching topics cover the Lexus hybrid models. The "Electric Cars" index is focused on electric cars. Among five other indexes relevant to hybrid cars are ones that cover their technology. By exploring the topics in indexes, users can identify indexes that most closely match their interests.

6. Sensemaking: Digital and Social

This chapter has introduced social indexing as a new form of social media. Social indexes address three sensemaking challenges: tracking core topics in our information diets, discovering information in frontier topics, and orienting ourselves to understand information in new subject areas. Social indexing remakes conventional indexes as *computational, trainable, social, and interconnected.* This approach follows the trajectory of emerging technologies for social media. It leverages the activities and knowledge of information communities, helping sensemakers to find both answers and the "right questions."

Acknowledgments

Thanks to my colleagues Eric Bier, Dorrit Billman, Dan Bobrow, Stuart Card, Ed Chi, Jeffrey Cooper, Markus Fromherz, Randy Gobbel, Lichan Hong, Bill Janssen, Joshua Lederberg, Lawrence Lee, Peter Pirolli, and Leila Takayama for their very helpful comments on this chapter. Special thanks to the Kiffets social-indexing team—Lance Good, Sanjay Mittal, Priti Mittal, Barbara Stefik, and Ryan Viglizzo—who have joined me in developing the social-indexing vision and making it a vibrant web-based reality. Thanks to PARC management for its support for a new venture during a difficult economic period.

Notes

1. In long-tail distributions the most popular information and media represent the head of the curve; consumption of items in this region dwarfs that of items farther down the tail (Anderson 2006, 1). Most of us, that is, consume the few items at the head, while our selections farther down the tail are more idiosyncratic.

2. Imagine a circle representing a central topic of interest surrounded with other topic circles of equal size. The combined area of these immediate frontier neighbors is six times the area of the central circle.

3. Larger granularities of topic are represented by books' organization into parts, chapters, and sections. My sampling revealed regularities here as well. The number of

chapters varied with the overall length of the book and the complexity of the subject matter. Where there were many chapters, they were often grouped into larger parts. Conversely, chapters were divided into sections and sometimes subsections. The number of hierarchical levels tended to increase with the length of the book, with a transition from two levels to three levels at around 500 pages, or 150,000 words.

4. For efficiency, our query generator employs best-first, anytime algorithms that attempt to visit the most likely parts of the search space first, and manage time and storage-space budgets to focus the search. Branches of the search process are pruned early if it can be determined that they cannot possibly yield candidates that will score better than queries that have already been generated. Because many candidates are eliminated after only a partial generation and partial evaluation, the reported candidates represent only the tip of the iceberg of the queries considered by the generator.

5. The structural complexity of a query is a measure that increases as a query becomes more elaborate, with more predicates, terms, and levels. By favoring low-complexity candidates the program follows the philosophy of Occam's razor, choosing the simplest queries that explain the data. Considerations of structural complexity are also helpful to avoid overfitting in the machine-learning process, especially when the training data are sparse.

6. Topic drift is a phenomenon that arises as news evolves. For example, one of our indexes about golf had a topic about Tiger Woods. It was originally trained on articles written when he was recuperating from a knee injury. Later, when he returned to competition, the query failed to pick up some new stories. Some additional training examples were provided, causing the system to adjust its topic models to accommodate the new stories.

7. In Wikipedia controversial stories are flagged and can be an interesting barometer of active debates. It may be possible to detect controversial articles by their pattern of vote cancellation.

8. Online news sites like Reddit characterize their different topic areas as "communities," but there are no membership requirements. At the time of this writing, the community structure in Reddit was almost identical to the topic structure in Digg. What seem to be needed in the next generation of tools are social indexes created by more specialized and dedicated communities. By comparison, the social processes that are active on Wikipedia for collaborative writing of articles seem more effective.

9. Some online sites (such as http://grou.ps) provide tools for creating social networks to share photos or collaborate. At present, these sites seem to be designed to help people maintain social connections, not to support sensemaking with evergreen indexes. Communities on these sites do not interlink to define information frontiers.

10. Such disagreement may also indicate an impending split of the community, as happens in the life cycles of scientific fields, churches, and political parties. An interesting design challenge arises from the tension between giving experts extra influence to keep the service responsive and limiting expert influence when the field or subject is shifting and the old guard is not keeping up.

11. Geographically organized information is becoming increasingly common on the web, especially for mobile services. Google Maps is one of the best-known examples. Another example is Yelp, an online collection of reviews of restaurants and other retail services organized by city and neighborhood.

12. Degree-of-interest functions are also used in collaborative filtering, where a person's preferences regarding a sample of media are matched against collective prefer-

ences in order to predict additional interests and qualify recommendations. The computation involves estimating a person's closeness to a group. Although a frontier degree-of-interest function can employ an estimate of distance within a social network, the degree of interest is based not on matching preferences but on explicit designations of neighborhood. Furthermore, the degree of interest for an article can be weighted depending on whether it matches one or more of the subtopics in the home index.

13. "Orienting" is not to be confused with a similar-sounding topic in search behavior, "orienteering." As described by Jaime Teevan and others, orienteering involves using prior and contextual information to narrow in on an information target. The searcher generally does not and cannot specify the complete information need at the beginning. The term "orienting" comes instead from the analysis of sensory systems, where there is typically some "alert" that causes an all-hands-on-deck cognitive response. This is the "orienting response." In our use of the term "orienting," we refer both to this point, when attention is drawn to relevant material, and to the providing of additional topical cues for understanding the meaning of the material.

References

Anderson, Chris. 2006. *The long tail: Why the future of business is selling less of more.* New York: Hyperion.

Brin, Sergey, and Lawrence Page. 1998. The anatomy of a large-scale hypertextual web search engine. 7th International Conference on World Wide Web, Brisbane, Australia, April 14–18.

Burt, Ronald D. 2004. Structural holes and good ideas. *American Journal of Sociology* 110, no. 2: 349–99.

Card, Stuart, Jock Mackinlay, and Ben Shneiderman. 1999. *Readings in information visualization: Using vision to think.* San Francisco: Morgan Kaufmann.

Chi, Edward H., Lichan Hong, Julie Heiser, and Stuart K. Card. 2004. eBooks with indexes that reorganize conceptually. Human Factors in Computing Systems Conference, Vienna, Austria, April 24–29.

Pirolli, Peter. 2007. *Information foraging theory: Adaptive interaction with information.* Oxford: Oxford University Press.

Simon, Herbert. 1971. Designing organizations for an information-rich world. In *Communications and the Public Interest,* ed. Martin Greenberger, 37–72. Baltimore: Johns Hopkins University Press.

Stefik, Mark. 1995. *Introduction to knowledge systems.* San Francisco: Morgan Kaufmann.

Stefik, Mark, and Barbara Stefik. 2004. *Breakthrough: Stories and strategies of radical innovation.* Cambridge, MA: MIT Press.