Social Indexing

Mark Stefik
Palo Alto Research Center
3333 Coyote Hill Road, Palo Alto, California 94304
Stefik@parc.com

ABSTRACT
This paper identifies three challenge problems for sensemaking: focusing on information in three tiers: core interests, information frontiers, and new subject areas. In addressing the challenges we take a fresh look at the old idea of indexes, recasting them as computational, trainable, social, and interconnected. The new *social indexes* leverage the activities and knowledge of information communities, helping sensemakers to find both answers and the "right questions."

Author Keywords
Sensemaking, social indexes, social media.

ACM Classification Keywords
H5.3. Group and Organization Interfaces, H.5 Information Interfaces and Presentation (e.g., HCI).

INTRODUCTION
Sensemaking is the process by which we go about understanding the world. "Digital sensemaking" is sensemaking intermediated by a digital information infrastructure, such as today's web and search engines.

Although digital sensemaking today is mostly a solitary activity, social media approaches are now emerging that may radically change the experience of digital sensemaking. Web search engines have greatly improved our ability to find information. However, tools for sensemaking have not nearly reached their potential. For most people sensemaking on the web is often frustrating and onerous, requiring them to wade through off-topic, and poorly-written web pages of questionable authority.

Even professional sensemakers experience failure and frustration with current sensemaking tools. Anyone following the interplay of information and politics in the intelligence community has noticed some very public and remarkable strategic failures. Over the past few years, these include the failure to appreciate the warning signs before the Yom Kippur War in Israel, the unanticipated collapse of the Soviet Union, and more recently the apparently false conclusion that there were weapons of mass destruction in Iraq. Intelligence analysts are the jet pilots of sensemaking, meaning that their professional challenges are extreme. Intelligence analysis involves requesting and collecting an immense amount of information, sorting through it to identify relevant pieces, and constructing and maintaining an ongoing understanding for tactical and strategic purposes. 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 theme is that besides finding the right answers it is important to find the right questions. The second theme is that there is more power available for sensemaking when we cast it as a social activity rather than as an individual one.

INFORMATION AND ATTENTION
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. How we consume information *individually*, however, is better understood in terms of "information diets." Our personal information diets describe how we allocate our time across different categories of information, including both popular and specialized topics.

Each of us has a personal information diet. The term information diet characterizes the information that we consume across different categories including topics in the news, professional interests, hobbies, and entertainment media. An information diet can be represented as a list of topics together with the corresponding allocations of our time or attention. The total allocations add up to one hundred percent of our available time.

Although the categories are different for each person, what we have in common is that our individual information diets do not align with the long tail popularity curve. For example, the top stories of the daily news may occupy a minor share of my daily information diet, even though they are popular for the general public. Except for teenagers, there are relatively few people for whom the top item in their personal news diet is the most popular news story;

CHI 2008, Sensemaking Workshop April 5–10, 2008, Florence, Italy.

their current favorite piece of music is the most popular tune, and so on, consistently following the dictates of popular taste.

Current search tools and news services are optimized to serve the head of the long tail. Unfortunately, information infrastructure that is optimized to serve us in the aggregate does not serve us very well as individuals.

THREE CHALLENGES FOR DIGITAL SENSEMAKING

Discovering information is difficult, especially when information sources are dynamic. For an information consumer, the challenge is to quickly discover information that is new, relevant and important. New information becomes available from many different sources.

Web search engines are not ideal for information discovery. They do not enable us easily to focus on a subject area or topic, and typing "What's new?" into a 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 algorithms like PageRank [3], which depend on inter-page 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.

Online news services are also far from ideal. They cover only broad popular topics. For specialized topics, a myriad of syndication feeds cover topical information from much farther down the tail. There are now hundreds of thousands of such sites on the web, and today's readers and aggregators do not sort effectively through the tide with an eye to quality and authority.

News aggregators, both professional and automatic, generally fail to group stories coherently by topic. Mainstream news services use very broad categories such as "business," "national," "international," "entertainment," and "sports." Some online news services offer key word retrieval of news articles, including email alert services. However, these search results provide few means for judging the reliability or credibility of articles and are prone to incoherently mixing and scattering ones on similar topics.

This brings us to our first challenge for sensemaking and information foraging: *better approaches for tracking or discovering information* for our core interests.

In his characterization of personal knowledge [8], Michael Polyani differentiates between proximal and distal knowledge. Proximal knowledge corresponds to knowledge that is "closer in." It is familiar and readily available to us. Distal knowledge is farther from our core interests. We may know people who are familiar with it, but we are not personally familiar with our distal information topics.

Information beyond the edges of our interests constitutes our "information frontiers." The frontiers in professional fields are topics from related and nearby fields. Information frontiers in community news are often news from neighboring communities. Information frontiers in business and technology reveal new developments that bring change and opportunity. Frontier explorations are both crucial to people interested in spotting new trends.

As a director at the Institute for the Future, Paul Saffo is in the business of analyzing technology and business futures. In an interview about their forward-looking process [9] 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."

Saffo's interests in the frontiers or edges of a field bring to mind Ronald Burt's ideas about structural holes in social structures [4]. People attend mainly to ideas from inside their group. This creates "holes" in the information flow between groups. Burt's hypothesis is that new ideas emerge from synthesis across groups. People who are connected across groups are more familiar with different ways of thinking. They have an advantage in detecting opportunities and synthesizing 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. Consequently, we need more help in allocating some of our scarce attention to scanning our information frontiers.

This brings us to our second challenge in information foraging, *better approaches for "prospecting" or mining our information frontiers*.

More often than we may realize, we also need to learn about topics that have not previously been of interest. For example, information on a new kind of appliance can become interesting when we first consider a purchase. Information specialties at work can become interesting when we need to substitute for a co-worker who is taking a leave. Learning about a new category of medical information can become urgent when someone develops a health problem and the family urgently needs to learn about treatments and services.

This brings us to our third challenge in information foraging, *better approaches to get "oriented" in an unfamiliar subject area*.

In summary, the three challenge problems for digital sensemaking address information in three tiers. The information discovery challenge focuses on core interests. The information prospecting challenge focuses on our

information frontiers. The information orientation challenge focuses on new subject areas.

INFORMATION DISCOVERY FOR CORE INTERESTS

In a typical scenario for the discovery challenge, sensemakers have access to a corpus of pre-existing materials with information on their core topics. New materials arrive from multiple sources. The new materials are not categorized by subtopic and may include information beyond the information diet. Levels of authoritativeness may vary. The challenge is to classify the new materials at fine grain by subtopic and to determine a quantified degree of interest for prioritizing articles and allocating attention.

An automatic approach for improving information discovery must address two key sub-problems: computing an evergreen index that sorts new information by topic, and determining a degree of interest for the relevant articles. Our approach for addressing information discovery is the subject of active research in our laboratory. The following discussion is based on our ongoing experience with prototype systems that we have built and are using to develop and test the approach.

Problems with Indexes

Various earlier approaches for automatic indexing have received research attention such as indexes based on concordances. A concordance is a listing of words and phrases found in a document together with their immediate contexts. A concordance is built by marching through a source to identify the terms and phrases that are present, noting the multiple pages on which they appear, and creating an alphabetized list of them. Concordance-like back-of-the-book indexes can be computed automatically, sometimes using linguistic techniques for phrase selection and normalization.

Concordance-based indexes fall short for the information discovery challenge because their articulation of subtopics is not informed by domain expertise or historical experience. Although concordances systematically list the phrases that appear in a document, they do not identify and carve material along the ontological and topical "joints" that are useful and used by people in a field. They fail to distinguish between the important and the trivial. For these reasons they are inadequate for the needs and purposes of information discovery.

To explain a new approach using indexes to support discovery, we first begin with the familiar example of a book index, and then show how to develop an approach for the web and dynamic news sources. A well-written text book comes with an index that embodies judgments of how people will use the information in the book. The index entries reflect an expert's articulation of the important topics and a list of pages in the book where each topic is discussed. Although a book index is a good starting point, an 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 document collections and online sources. What is needed is an automatic approach to extend an index to new material.

Generating Index Patterns

In our laboratory we have developed an approach to this problem called *index extrapolation.* Index extrapolation takes an existing index and uses it to bootstrap an evergreen index. The existing index serves as training set for a machine learning system that updates the index as new material arrives.

Since index extrapolation has not been previously reported, it may be of interest to explain how it works. Our approach to index extrapolation uses a hierarchical generate-and-test algorithm [10]. The broad steps of index extrapolation are as follows. For each subtopic in the index, the index extrapolation program analyzes the pages that it cites. It selects a subset of "seed" words on these pages that seem characteristic of the subtopic. These characteristic words are those whose frequencies on the cited pages are substantially higher than their frequencies on the other pages. Other words may be included as seeds when they are part of the subtopic's label or are near a label word in the cited text.

Index Entry	Generated Pattern	Meaning
Afghanistan:: Soviet conflict with	[afghanistan with $Russia]	The term Afghanistan and the term "with" and the library pattern for "Russia".
aflatoxin	aflatoxin	The term "aflatoxin".
African swine fever	{african swine fever}	The ngram "African swine fever"
Against the Grain (Yeltsin)	[yeltsin {against the grain}]	The term "Yeltsin" and the ngram "against the grain".
Black Death	(bubonic {black death})	The term "bubonic" or the ngram "Black Death".
Bonfire project	[bonfire (project program)]	The term "bonfire" and either the term "project" or the term "program."
Biodefense: superterrorism and	[destruction mass (superterrorism terrorist)]	The term "destruction" and the term "mass" and either the term "superterrorism" or the term "terrorist"

Table 1. Sample index entries from [1] showing the patterns generated for each subtopic.

The index extrapolation system then begins a systematic, combinatorial process to generate candidate patterns in a finite-state pattern language. The patterns express subtopic recognition constraints in terms of four kinds of predicates: conjunctions, disjunctions, sequences, and ngrams (sequences of words with no other words in between). The patterns include single-level expressions over the seed words and multi-level expressions that include other

predicates as arguments. For example, a pattern might require that a particular seed word appear together with an ngram of three other words or a disjunction of two words. Tens or hundreds of thousands of candidate patterns or more are generated. The index extrapolator matches the candidates against the known corpus. Candidate patterns are rated according to how well they predict the actual pages cited for their subtopics in the training index. A candidate pattern performs perfectly when it matches all of the cited pages and none of the other pages. In other words, the ideal matching performance has no false positives and no false negatives. To choose a top pattern when multiple candidates exhibit perfect performance, the evaluator also considers structural complexity and term overlap with the index label.

The result of the machine learning phase is a pattern generated for every subtopic in the index. For example, in a test run using a book by a defector from the Russian intelligence community, the 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," "pandemic," and others. About a thousand candidate patterns were automatically generated and reported using the seed words. The top rated pattern required that a page contain either the word "bubonic" or the ngram "black death". An alternative candidate pattern 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." Yet another candidate required that a page include the word "plague," either the word "pandemic" or "rare", and either the word "yersinia" or "bubonic." The candidate pattern selected as the top 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. See Table 1 for more examples.

Keeping an Index Evergreen

The index extrapolation system keeps an index evergreen to new, arriving information. New pages are classified by subtopic by matching them against the patterns. When a new page matches a pattern, 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 patterns are generated automatically by machine learning rather than manually, and that the patterns are organized in a hierarchical topical index. The patterns for subtopics deep in a topic tree are more specific and tend to be more complex than patterns higher up.

As a corpus grows, new pages may show up that should be included under a topic but which are not matched by the pattern. When such pages are identified by a human editor or a voting process, they are logged as new, positive training examples. 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 improves the quality of the index.

Determining Degree of Interest

Index extrapolation technology addresses the first sub-problem of the information discovery challenge: maintaining an evergreen index. We now turn to the second sub-problem: 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.

Digg pioneered a social media approach to rating and ranking news stories. It is 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 news services, and also to vote on them. Digg also engages a social network of its readers. Members can subscribe to the stories that a friend or thought leader "diggs". It maintains a list of current stories prioritized by their votes. As a story gets positive votes it rises on the list. As a story 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 down and disappear.

This voting approach to ranking stories involves 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 can also digg the story causing it to shoot upwards. 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 c|net report [7] 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 discovery. Some typical criticisms of Digg are that it is too focused on technology 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 constant tug-of-war on votes for and against a controversial article. Because the polarized votes cancel each other, a controversial article would not rise up the popularity ranking. This brings to mind the cautionary advice about conversations with strangers: "Don't talk about politics or religion."

What kinds of articles appear on Digg? Even as the U.S. presidential election year approaches, there are no Digg categories for politics or religion. The category mix is more weighed towards technology than general news services. At the time of writing, Digg had forty nine classifications for articles, sorted under seven 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. Taking a sample of the Digg front page on the day I wrote this, there were fifteen articles. Eight of these were about the technology industry, including one about Digg and several about the web. Three articles 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 (about the network, operating systems, and video games) presumably because they are important to the Digg community. Even in a specialized Digg topic area such as International and Business, the articles are far from the mainstream reflecting a selection that is 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 having similar interests.

In summary, current systems for rating news socially suffer from several problems. The dominance of cliques in promoting articles is a variant of the "tyranny of the minority." The suppression of controversial topics by vote cancelling 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.

Augmented Information Communities

Organizing users into communities addresses the tyranny of the majority issue by making it possible for small groups and communities to explore core topics of interest to them. Each community has its own index, covering topics in its subject area. Within a subject area, communities could pursue their particular niches in the long tail, rating materials according to their own values. In a technical subject area, professional groups may focus on advanced materials while amateur groups focus on introductory ones.

Most users would belong to multiple communities, for each of the core topics in their personal information diets. For example, a user may belong to one or more communities on professional topics, a sports community reflecting a local team, a political news community with people of similar interests, a hobby-related community, and so on. Different communities may cover similar topics. For example, there may be "red," "blue," and "green" communities for political news and perspectives with different Republican, Democratic, and environmental slants.

Following this approach, Figure 1 shows a mockup design for an index display for the "iPhone" subtopic. The tabs at the top of the display show four communities from a larger set, each of which has an index about iPhones. The largest community, "Tech News," has 26,643 members. In the Tech News community, "iPhone" is a subtopic of "Cell Phones," which is a subtopic of "Sci/Tech." Subtopics of iPhone include "Applications," "New Models," and "Reviews." The next four communities are "Developers," "Phreaks," and "Consumerz". By clicking on the corresponding community tags, a user would switch to a different community with its own index and ranking of articles.

The degree of interest computed for articles governs their display in the index. For example, Figure 1 shows articles under the "Applications" subtopic in a larger font than the other two subtopics based on an evaluation that they are more popular. Within a subtopic column, the most popular (and presumably important) articles come first. The system allocates different amounts of space to articles depending on their ratings. For example, in the "New Models" column, the first two articles are ranked as the most important. They are allocated more space, allowing for the display of a graphic, a title, information about the source of the article, and an abstract or beginning paragraph. The third article in the column is allocated less space, includes no graphic, and has a smaller font. At the bottom of the column, three articles are represented only by their sources. The remaining articles are not shown at all, except for the "more" link for seeing additional articles. This approach

allocates an amount of space for each index citation according to a degree-of-interest function [6]. Given the space allocation, it then determines the appropriate amount, kinds and format of information.

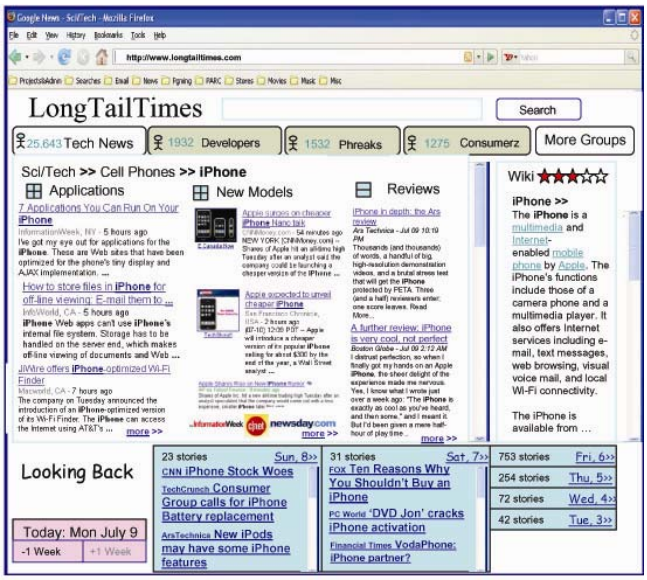

Figure 1. Mockup of a web-based user interface showing multiple communities and part of an index for one.

In summary, it is useful to divide a population into communities of interest. The placement and space allocated to displaying articles can be governed by the community's voting practices.

It is also useful to provide transparency across communities when they are interested in related topics. Communities can become 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 visibility into the topics and discussions of other communities, a discovery system can have a broadening influence by showing members other views of the world.

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 overly subject to the tyranny of cliques? How do communities keep from getting too self-focused and narrow?

The genesis of an online community could begin when a founding individual decides to pursue some interests by starting a private index. The founder defines an initial set of online sources, such as new feeds, web sites, 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 patterns 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 can be admitted at different levels. For example, "expert members" might be defined as a set of thought leaders in a community. An initial set of experts could be identified. New members could gain expert status on the basis of various social actions, using familiar social qualification mechanisms involving referral, voting, recommendations, and so on. Votes by experts would have more influence in ranking articles than for regular community members. Experts could also have expanded roles in maintaining the structure of the index by occasionally creating and editing topics.

Another category of users might be "harbingers." A harbinger is a community member with an extensive voting record who tends to be early, accurate, and prolific in identifying articles that the community ultimately ranks highly. Whereas "experts" are appointed or elected, harbingers could be discovered automatically. They are accurate predictors of a community's interests and values. Harbingers are identified by tracking their submissions and votes over time. As they are qualified, their votes are given more weight than the votes of regular members. If harbingers or experts have a streak of voting which is out of alignment with the community, their influence could automatically wane. Having expert or harbinger status in one community does not give one similar status in a separate community.

In summary, information communities could have different levels of memberships. Visitors could use the index and read the recommended information. Unlike the regular members, visitors would not have any voting influence in rating articles. As reliable predictors, harbingers would have amplified influence in their votes. Experts would be qualified by various social processes and would also have amplified influence. Experts would also have a role in adding to and restructuring the index.

The Few, the Many, and the Machines

This approach for addressing the information discovery challenge relies on three sources of power. One source of power is the hard work of the few. The "few" are the experts who use their knowledge to create and maintain a topical index. The second source of power is the light work of the many. The "many" are the people who identify and vote on disputed citations, influencing the training sets for tuning the patterns. The third source of power is the tireless work of the machines. This refers to the index extrapolation algorithms that automatically match the patterns 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 generate patterns in a combinatorial search space and evaluate them to faithfully model the subtopics. Tireless by nature, computers can be massively deployed as needed to meet the scale of the information and usage.

These three sources of power are synergistic and fundamental to the design of social media.

PROSPECTING FOR FRONTIER INFORMATION

Prospecting refers to tracking materials in information frontiers, that is, in nearby subject areas. Frontier information is typically less important than core topics. At the same time, the subject matter along a frontier is typically larger than the subject matter in a central core. Furthermore, the level of expertise of a sensemaker is lower for frontiers than for core subject areas, both for identifying good sources and understanding the topic structure. Although it is tempting to ignore the frontier, there is a risk. 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 in the information discovery challenge, the value proposition of prospecting is better attention management. There are three sub-problems. The first sub-problem 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.

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 granularity of individuals, a social network expresses relationships among people with common interests. At the granularity of communities, a social network 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, where members focus their attention on core topics in a subject area. Neighboring communities represent other fields or other groups. The technology for prospecting a frontier is intended to selectively provide a "vision advantage" that can be used for synthesizing new ideas and spotting trends.

When the leaders of a home community want to be made aware of relevant articles that another augmented community finds interesting, they can designate it as a frontier neighbor. In a simple approach, candidates for neighbors could be found manually by searching a directory of communities. In a more sophisticated approach, the multi-community indexing system could suggest candidate neighbors using similarity measures that detect overlap of interesting sources and articles in pairs of communities.

For 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 is another 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, suppose that there is a social index for the city of Palo Alto, California where I work. Immediate neighbor cities of Palo Alto include Mountain View, Los Altos, Stanford University, Menlo Park, and East Palo Alto. For a medical example, a community interested in traditional Chinese medicine may focus on traditional acupuncture and herbology. The community would be distinct from the myriad of "New Age" medical approaches in the west, although it might choose to connect to such communities or Ayurvedic (Indian) medicine as its frontier. Networks of augmented communities could 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.

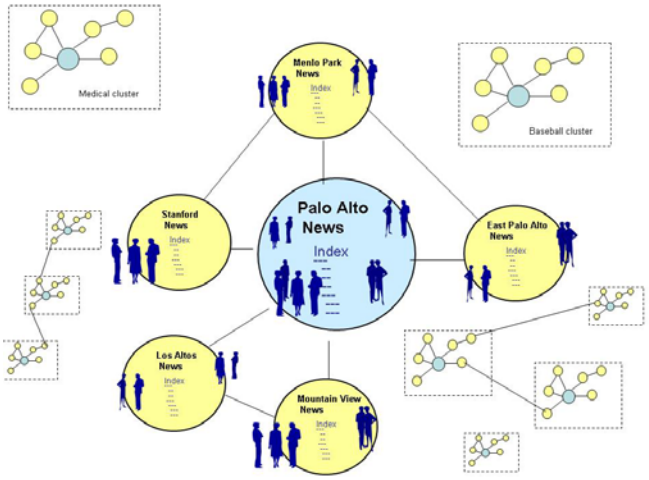

Figure 2. Social network of augmented communities.

In summary, each augmented information community has its own index, its own members, its own information sources, and its own ratings. By explicitly connecting to each other, augmented communities can define and gain vision into their information frontiers.

Relating Frontier Information

The third sub-problem is to relate the frontier articles to home topics. Few articles from the frontier will have universal interest across a home community. One approach is to use the home index to automatically re-classify articles by the subtopics that they match in the home index. In this way, articles are 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 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.

ORIENTATION TO NEW SUBJECT AREAS

Our third sensemaking and information foraging challenge is orienting to information that is completely separate from a personal information diet. Orientation refers to a process of getting familiar with a subject area, such as by learning about its topical structure, main results and best references in order to answer questions important to the sensemaker. The orientation challenge arises whenever we need to learn about something completely new.

The orientation 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 the orientation challenge, we consider again fundamental properties of a good index. An index is a layered organization of topics. A good index embodies expert judgments about how people in the community want to use 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 the subject 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 provided specifically for orientation. Following this line of thought, *the challenge of orientation is largely addressed by providing the sensemaker with a good index.*

Figure 3 is a screen shot from a prototype index extrapolation system built in our laboratory. The user interface shows two panes: a reading pane and an index pane. The reading pane is used for navigation among pages in the book. A sensemaker can request a guide to more *information on the topics of the page being read*. This causes a specialized index or "guide" to be generated and displayed on the right. The guide page shows the relevant subset of the book index, limited to the subtopics represented in the text of the reading page. A simple way to identify the relevant topics is to match the generated patterns for all of the subtopics in the index, creating a lexicographic list of the subtopics whose patterns match the page. Alternatively, one can compute a "scent index" given a query [5], basing the subset of selected subtopics on a match to a query.

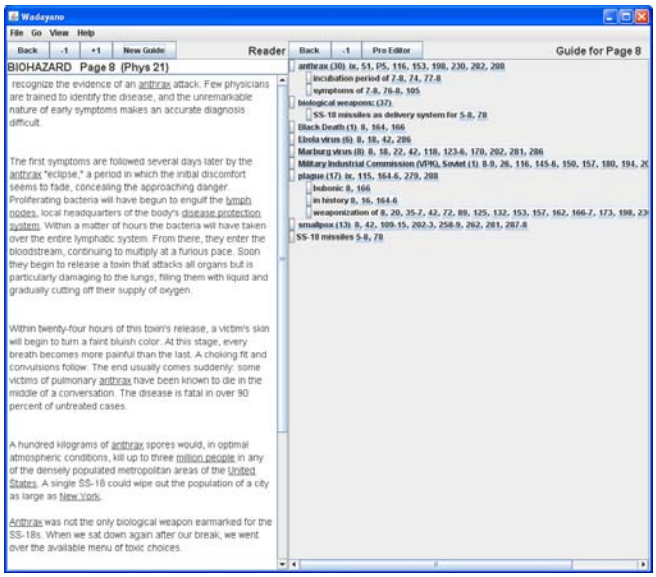

Figure 3. Screen shot from an index extrapolation system

In an orientation scenario, suppose that a sensemaker has identified a page or so of information on a topic, perhaps after issuing a query. The sensemaker considers the page useful and wants to become more familiar with the subject area. However, the sensemaker does not yet know which community index to use.

The pattern matching capabilities of an active index lend themselves to a novel way of supporting orientation. It can match *all of the topic patterns from tens or hundreds of thousands of indexes against a sample page i*n a small fraction of a second. In this way, the system collects all of the potentially relevant community indexes. The next step is to rate the competing indexes, choosing those that have the most relevant subtopics and the most relevant articles. The system can then use the same degree-of-interest concepts to create a display similar to Figure 1. The tabs would show the competing augmented communities in ranked order. The sensemaker could then browse through articles from each community. In this way, a sensemaker can discover both a guiding index and an augmented community of like-minded people. The index entries provide an introduction to the important questions of the subject area.

Using an interface similar in function to that in Figure 3, the user can alternately explore new topics or further references on a topic. This enables a form of dual search. One part of the search is across articles on a topic. At any time, a sensemaker reading a page can get a subset index for exploring topics on the page more deeply. The second part of the search is a search across the topics that make up community's conceptual organization of the subject area. From any topic, one can access pages or articles on the

topic. Pursuing the dual search, the sensemaker becomes oriented to the new subject area.

In summary, a system for the orientation challenge exploits a directory of augmented communities. The sensemaking system helps select an augmented community and computes a subset index.

CONCLUSION

This paper introduced social indexing as a new form of social media. Social indexing addresses three challenge problems for sensemaking for information in three tiers: core topics, information frontiers, and new subject areas. Social indexes leverage 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, Barbara Stefik, and Leila Takayama for their very helpful comments on this chapter. A longer version of this paper will appear in [2] with the title "We Digital Sensemakers.".

REFERENCES

1. Alibek, K. *Biohazard*. New York: Dell Publishing, 1999.
2. Bartscherer, T. & Coover, R. *Switching Codes.* Chicago: University of Chicago Press (in press)
3. Brin, S. and Page, L. The anatomy of a large-scale hypertextual Web search engine. *Seventh International Conference on World Wide Web*. (1998), 107-117.
4. Burt, R.D. Structural holes and good ideas. *American Journal of Sociology* (2004), 110:2. 349-99.
5. Chi, E., Hong, L., Heiser, J., & Card, S. eBooks with indexes that reorganize conceptually. SIGCHI Conference on Human Factors in Computing Systems. (2004), 1223-1226.
6. Furnas, G.W. Generalized fisheye views. *SIGCHI conference on Human Factors in Computing Systems* (1986), 16-23.
7. Mills, E. The big Digg rig. c|net News.com (2006) http://www.news.com/2100-1025_3-6140293.html
8. Polyani, M. *Personal Knowledge: Towards a Post-Critical Philosophy*, Chicago: University of Chicago Press (1998).
9. Stefik, M. & Stefik, B. *Breakthrough: Stories and Strategies of Radical Innovation.* Cambridge: The MIT Press (2004), 167-8.
10. Stefik, M. Introduction to Knowledge Systems. San Francisco, Morgan Kaufmann Publishers (1995), 259-80.