I-DIAG: From Community Discussion to Knowledge Distillation

Mark S. Ackerman[a,b], **Kurt DeMaagd**[b,c], **Stephen Cotterill**[a], and **Anne Swenson**[a,b]

[a] Electrical Engineering and Computer Science, University of Michigan
[b] School of Information, University of Michigan
[c] Haas School of Management, University of Michigan

{ackerm, demaagdk, scotteri, aswenson}@umich.edu

Abstract

I-DIAG is an attempt to understand how to take the collective discussions of a large group of people and distill the messages and documents into more succinct, durable knowledge. I-DIAG is a distributed environment that includes two separate applications, CyberForum and Consolidate. The goals of the project, the architecture of I-DIAG, and the two applications are described here.

Introduction

Imagine the following scenario: The president of a large public university in the US asks a blue-ribbon panel of his highly regarded faculty to reflect upon the future of their university. The president wants to keep their university not only in the forefront of similar universities but also in front of basic societal pressures and opportunities. However, the faculty are also admonished to consider the various often-overlooked stakeholders – the university's staff, undergraduate students, graduate students, alumni, non-tenured instructors, state legislature members, and local community residents. A large US state university may have several thousand faculty members, and the various concerned stakeholders might include 50 thousand or more people. Of course, the faculty committee could do as a typical blue-ribbon panel often does, going into their respective rooms to inscribe their already acquired expertise. But if they wished, how might they reach out to these stakeholders, include their perhaps divergent opinions, and search for new and interesting opinions and options?

We know that the Net is good at providing forums for large groups (> 10^5 people) to gather, discuss, and trade ideas. Within a corporate setting, this can be used for brainstorming, new produce ideas, quality circles, and the like. Governments, institutions, and universities can discuss such issues as organizational change and future plans in order to come to a "shared mind".

Yet all too often problems arise in these attempts. People do not come to the site, or do not stay on topic. More importantly, once use has finished (either by deadline or by neglect), the site is often a bramble of ideas and topics, too large and unwieldy for its information to be successfully reused.

Our system, I-DIAG[1], investigates how to garner and then distill this valuable community knowledge. It is part of a larger project to investigate how to maintain and reuse informal information within organizational and Internet-scale settings.

This paper reports on I-DIAG as a work in progress. The paper is arranged as follows: We begin with a description of the research problems under consideration, and follow that with a brief overview of the relevant literatures. We then discuss the architecture of I-DIAG as well as provide a description of the various components of I-DIAG. (I-DIAG consists of a number of applications and distributed services.) We conclude with future work and directions.

Research Overview

We created I-DIAG to consider several general research problems as well as provide a concrete application with which to examine these problems. Overall, we are investigating:

- ❑ New models for refinement and distillation. Our primary interest is in finding social and technical mechanisms to facilitate the distillation of knowledge from large amounts of informal information, such as bulletin-board messages, chat messages, e-mail, or quickly written brief documents. Our argument below

[1] The main quad of the University of Michigan campus is called the Diag. I-DIAG is also short for Interactive Diagenesis. Diagensis is "the recombination or rearrangement of constituents (as of a chemical or mineral) resulting in a new product, or the conversion (as by compaction) of sediment into rock (Webster's 1986)".

is that previous mechanisms have failed because of the social barriers. Accordingly, our emphasis is less on the technical mechanisms for doing textual summarization or knowledge elicitation than on finding social models with augmentative technical mechanisms to foster the creation of material and then "boiling down" of that material into something that will be subsequently useful to others.

- ❑ These "boiled down" repositories are the distilled and refined versions of many people's thoughts about a subject, mostly likely specific to a particular socio-technical environment. We are also investigating mechanisms to foster the sustainability of this distilled repository over time.
- ❑ In any social space, mechanisms must exist to foster social regulation and sustainability over time (as in Ackerman and Palen 1996). While social regulation can have pejorative connotations for computer people, some amount is necessary to continue any collectivity's activities. It seems as though there are always problem or abusive users in online spaces. We also wish to prevent or ameliorate unproductive or hateful exchanges. As we will see, the duration for I-DIAG is very short. Nonetheless, there are still social regulation and maintenance issues to be resolved; indeed, some may be exacerbated by use assumed to be brief. Through I-DIAG we are investigating user and collaborative interface mechanisms to quickly move users into an understanding of the system and its uses, enable productive exchanges, and control potentially unruly users and problematic exchanges.
- ❑ Since we hope that use is rapid and the corpus of information is constructed very quickly, we are investigating interface mechanisms to allow users to return to the space and understand what is new quickly and effectively. We hope to produce interface guidelines for these types of spaces.
- ❑ Overall, we see ourselves as investigating new forms of knowledge management. I-DIAG forms an interactive or dynamic "book", where the corpus is constructed iteratively and collaboratively by people with different opinions, types of expertise, and varieties of experience and viewpoints. This "book" is a living document – not only is it constructed by people in terms of their own interests and knowledge, but it can be maintained over time in the same manner.

Our major goal, then, is to understand how to iteratively construct a refined knowledge repository (probably less than completely formalized but more distilled than raw messages). To do so, we must necessarily also investigate what technical *and* social mechanisms we need for sustainability, social regulation and maintenance, navigation and return, and interface metaphors.

In order to examine these broad issues, we have created a particular problem scenario and the computational system to solve it. The scenario in the introduction describes most of the problem we are addressing. It is a "brainstorming system", a system in which people can come together to offer ideas and debate them. Figures 2-4 show the specific testbed we have created to investigate these issues. A few points should be noted about the application and the environment:

- ❑ In keeping with the Internet philosophy of utilizing thousands of eyeballs, I-DIAG attempts to harness small amounts of time from thousands of users. Motivations for using the system come from everyday activity. We hope to have some small number of core users, who will be key contributors, but we expect small contributions from a much larger number of people. At the end, we expect only a handful of people to distill the material.
- ❑ In our standard scenario of use, we are assuming the site will be used actively for a brief period of time – two weeks in our current plans. This allows people to have a healthy and vigorous discussion on specific topics, and then the site can close down before the topic becomes obsolete or stale. It also provides us a time to start mining the discussion as a final product – namely the final report and/or a distilled, concise web site of responses and ideas.
- ❑ I-DIAG, accordingly, has three sets of users. The first user group consists of the people entering their comments and discussing appropriate topics. The second user group consists of the moderators, editors, and wizards that control the interactive discussion. The final set consists of the people distilling the archived materials, either for an external report or to create a more concise site.
- ❑ The precise outcome of any given I-DIAG installation may not be known in advance. Some instantiations may wish a linear book as their outcome. The distillation process for that would likely be different when one wishes a concise site as the outcome. In addition, the scope of the distillation might vary – some sites may wish to include every point of view and every significant issue; other sites may wish to merely keep subdiscussions or interesting points.

I-DIAG, then, is an attempt to reconsider knowledge management and knowledge communities. It attempts to create incentives for use and reuse by differing groups of people, all of whom iteratively construct the space and the knowledge through their activities.

Relevant Literatures and Related Systems

Several diverse literatures bring appropriate insights and prior work.

Of direct relevance here are a number of approaches to distillation and summarization. In an older Education literature, one can find descriptions of "advanced organizers," organization tools for structuring educational lessons or materials (Jonassen, Beissner, and Yacci 1993). Although over time, the term came to be known as a technique for textual or oral materials (similar to foreshadowing), originally these included visual organizers. These visual organizers included timelines, web of relationships, trees of concepts, and the like. Many of the visual interfaces are directly relevant to our efforts to provide organization tools to users; however, these visual interfaces, we feel, are only part of what is needed.

Similar in intent to the literature on visual organizers is an important paper on incremental formalization in hypertext (Shipman and Marshall 1999, Shipman and McCall 1999, Shipman and McCall 1994). Visual organizers allow one to slowly increase the amount of organization in one's material by presenting more conceptually-oriented views on that material. This idea has been generalized in Shipman's work. These papers argue is that one should consider how to allow incremental formalization over time: Users may wish to enter free text initially and slowly increase the level of organization and formalisms in their material. Incremental formalization is critical to how I-DIAG works.

As well, I-DIAG uses techniques derived from and similar to text summarization. Text summarization (e.g., Radev and Hovy 1999) attempts to consolidate large documents or sets of documents into abstracts or shorter documents. Many of the techniques are relevant to I-DIAG, but again these techniques are only part of what is needed.

I-DIAG is related to a number of different Computer Supported Cooperative Work (CSCW) systems (also called collaborative systems here). I-DIAG is obviously an e-community system. Largely studied for their social effects (e.g., Sproull and Kiesler 1991, Wenger 1998), these systems do not have a large technical research literature. (A general overview of deployed technology, however, can be found in Preece 2000.)

I-DIAG also has similarities to a variety of brainstorming systems that have been investigated over the years. Generally, most such systems have been deployed and studied within face-to-face and distributed work meetings (e.g., Nunamaker et al. 1991, Streitz et al. 1994). A number of studies have shown that the use brainstorming systems provides more ideas and more creative insight to a problem (Dennis et al. 1999). However, since the use of these systems has been limited to single-session meetings, little has been studied about the social structures of use over time, or the technology and human-computer interface mechanisms required to support that use over time.

One large-scale brainstorming system reported in the research literature was the White House 's Open Meeting on the National Performance Review (Hurwitz and Mallery 1995). Using the system, users "discussed, evaluated, and critiqued recommendations by linking their comments to points in the evolving policy hypertext."

As well, the evolving discussions in I-DIAG could serve as a rudimentary design or decision rationale system (Conklin and Begeman 1988, MacLean et al. 1990, Moran and Carroll 1996). In a decision rationale, users follow an ontology with an implicit social process in order to create a coherent, well-structured argument that can be viewed by others at a later time. The goal is to help future readers understand that decision and perhaps reuse portions of the rationale in their own decision processes. However, as Grudin points out in Moran and Carroll (1996), users must do considerable upfront work for an unclear future payoff, and thus, users will be reluctant to actually go to the extra work to construct the rationale argument. I-DIAG attempts to provide suitable incentives for all of the users of the system by separating the argumentation from the distillation.

Finally, in our own earlier work, we examined collaborative systems for the distillation process (Ackerman and McDonald 1996). The Collaborative Refinery (Co-Refinery) system supported a four-step process (although the steps could be done in any order). Collecting was the act of obtaining material for a collection, and culling was removing superfluous or redundant material from the collection. More importantly, Co-Refinery supported organizing and distilling the materials. I-DIAG takes its beginning point from Co-Refinery and its mechanisms.

In summary, considerable work has been done on creating, fostering, and governing e-communities. Systems have also been created to foster and support brainstorming and decision rationale on-line. However, there has been little work, to date, on distilling informal information, especially group brainstorming results.

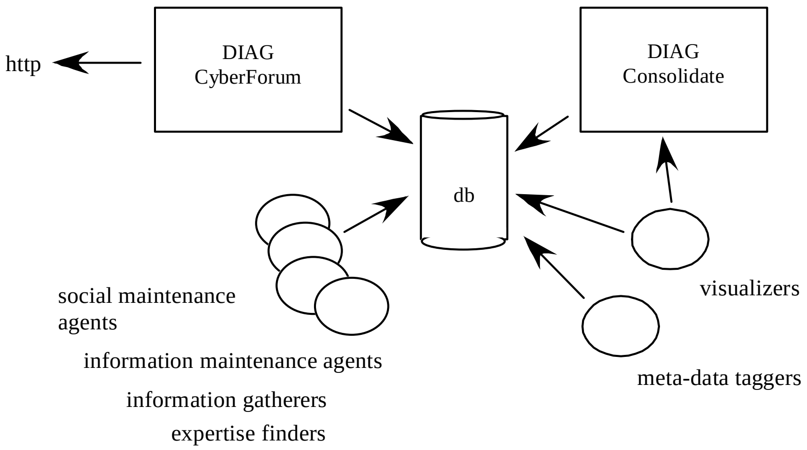


Figure 1: I-DIAG architecture

Architecture and Services

Differing users and their tasks suggested multiple applications, rather than trying to do everything in one Web-based application. For the discussion portion of I-DIAG, the interface requirements are relatively low. A Web-based interface could handle those requirements, and so we could consider customizing one of many Web-based discussion systems. On the other hand, there are substantial interface requirements for interactively handling sense-making, collaborative, and ad-hoc representations of complex intellectual spaces. Web-based interfaces would likely be marginal.

Therefore, we constructed I-DIAG instead as an environment into which new applications and auxiliary agents can easily be added. The architecture allows a gradation between user-controlled applications and autonomous agents. The architecture is shown in Figure 1. As many Web-based applications have, I-DIAG has a database at its core. For I-DIAG, the database stores largely hypermedia objects as well as meta-data. Applications (discussed immediately below) and agents feed to and from the database. As new services are developed, they can be placed into the architecture easily. We expect some of these services and applications to consist of relatively standard software projects; others will consist of research prototypes.

CyberForum

The front-end, discussion service is called I-DIAG/CyberForum (CyberForum). CyberForum is a typical Web discussion site. It is an application built upon the open-source Everything2 engine (essentially the same as that used by the Slashdot site). Everything2 provides CyberForum with the capability for message creation, editing, and storing. Support is provided for constructing displays, linking, and threading.

This application is absolutely essential to solving the scenario problem described in the introduction. While the effort of constructing the CyberForum application was relatively straightforward engineering, it did require substantial effort. The raw engine provides only the basic underlying services; without additional programming, little could be done. Figure 2 shows the CyberForum home page.

In addition to the basic application engineering, several research problems had to be addressed. As mentioned above, at a social level, we have had to consider additional collaborative mechanisms to facilitate social interaction and regulation. Because CyberForum is intended for relatively short-term use – a few weeks or a month for a particular site – the system has had to be optimized not only for performance efficiency (as does any Web application) but also for *social maintenance*. Social maintenance includes how to motivate users to come to and continue to participate at the site (*social facilitation*) as well as how to deal with problem users (*social regulation*). To support these requirements, we have added:

- ❑ Facilities to allow people to easy come in and out of discussions. In order for users to return to the site over time, it is important for them to be able to easily determine the current state of discussions as well as see what is new. We are providing a series of visualizations to help users orient themselves.
- ❑ User facilities to see what messages someone has posted. This not only provides a motivation for users to post, it also allows some pre-processing for later

distillation. Moderators can highlight interesting posts for other users. Moreover, they can annotate, discuss, or merely note interesting posts for later examination.

- Summaries to close problematic discussions. For example, one may wish to summarize the perennial and inconclusive argument about whether people should use Macs or PCs, closing off I-DIAG to this off-topic discussion. More importantly, it may be important to summarize and close off discussion of socially divisive arguments (e.g., affirmative action, the place of minorities in the institution under discussion). These summaries can provide a visual consolidation with further discussion allowed, a closing-off of further discussion, or a conclusion to an extended discussion. Figure 3 shows one summary, as well as a message thread.

We expect to add additional services to support the social requirements as we use CyberForum in limited field tests. Recently, we have begun to make our rating mechanism more flexible, especially with regard to the visual indicators for a message's rating by other users.

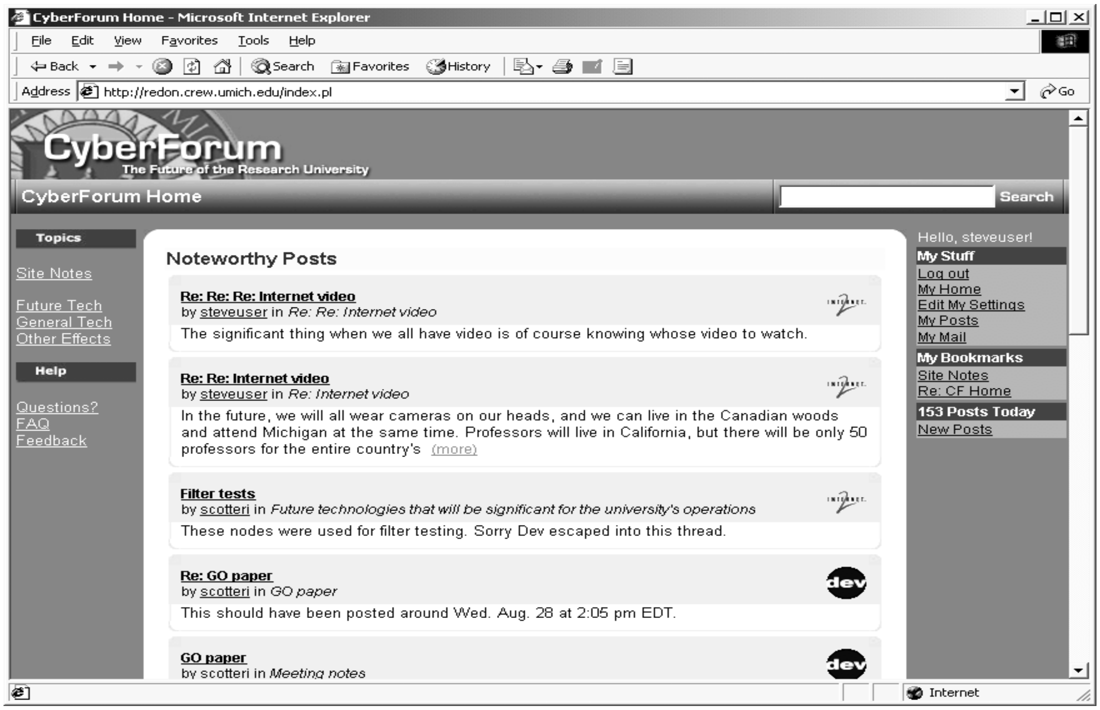
At a technical level, we had to add three major additional facilities to the Everything2 engine in order to create our computational architecture. In order to have external agents, we added a SOAP interface. Everything2 out of the box does not support communication with external programs. This facility gives us many additional capabilities. For example, we could more easily change the base rating system by using an external agent to monitor discrepancies.

To facilitate the social processes around editing and moderating of messages, we added a base layer of process support as well as a facility that allows the system to be easily reconfigured into a number of social formations. For example, we can easily switch from a free-flowing discussion to a moderated discussion (with a single moderator/editor) to a discussion moderated by an editorial board (where there must be a majority vote in favor). These latter facilities are critical to our efforts at social regulation and maintenance.

Consolidate

The second major application in the I-DIAG environment is I-DIAG/Consolidate (Consolidate). Consolidate, in our scenario of use, will be used by experts to consolidate and distill the messages and organization of the site once people have finished with CyberForum. Consolidate consists of an extremely flexible core system that ties together extensible views, a query service, and visualizations of the information (in this case, messages, threads, topics, and people) and its structure

Consolidate provides for collaborative use through shared views. The data for these shared views can be handled through a variety of replication engines; currently



http://redon.crew.umich.edu/index.pl

CyberForum Home

Topics

Site Notes

Future Tech
General Tech
Other Effects

Help

Questions?
FAQ
Feedback

Noteworthy Posts

Re: Re: Re: Internet video
by steveuser in *Re: Re: Internet video*
The significant thing when we all have video is of course knowing whose video to watch.

Re: Re: Internet video
by steveuser in *Re: Internet video*
In the future, we will all wear cameras on our heads, and we can live in the Canadian woods and attend Michigan at the same time. Professors will live in California, but there will be only 50 professors for the entire country's (more)

Filter tests
by scotteri in *Future technologies that will be significant for the university's operations*
These nodes were used for filter testing. Sorry Dev escaped into this thread.

Re: GO paper
by scotteri in *GO paper*
This should have been posted around Wed. Aug. 28 at 2:05 pm EDT.

GO paper
by scotteri in *Meeting notes*

Hello, steveuser!

My Stuff
Log out
My Home
Edit My Settings
My Posts
My Mail

My Bookmarks
Site Notes
Re: CF Home

153 Posts Today
New Posts

Figure 2: CyberForum home page

a simple replication scheme is supported. Through the shared views, multiple editors can discuss and consolidate differing organizations of the raw information. Multiple messages, as well as additional information (e.g., editor notes, links to external references), can be consolidated into summary nodes. Figure 4 shows an outline view of a topic; the icon in the lower right corner (which is normally in red) signals that this is a view shared with others.

In Consolidate, editors can place messages into multiple topics or even rearrange the topics themselves. While Everything2 and hence CyberForum-requires that all messages have only one parent, Consolidate does not. This is particularly important for knowledge distillation. Nodes can clearly be about multiple topics. In addition, editors may wish to keep their own lists of interesting nodes, nodes by certain people, and other kinds of working lists.

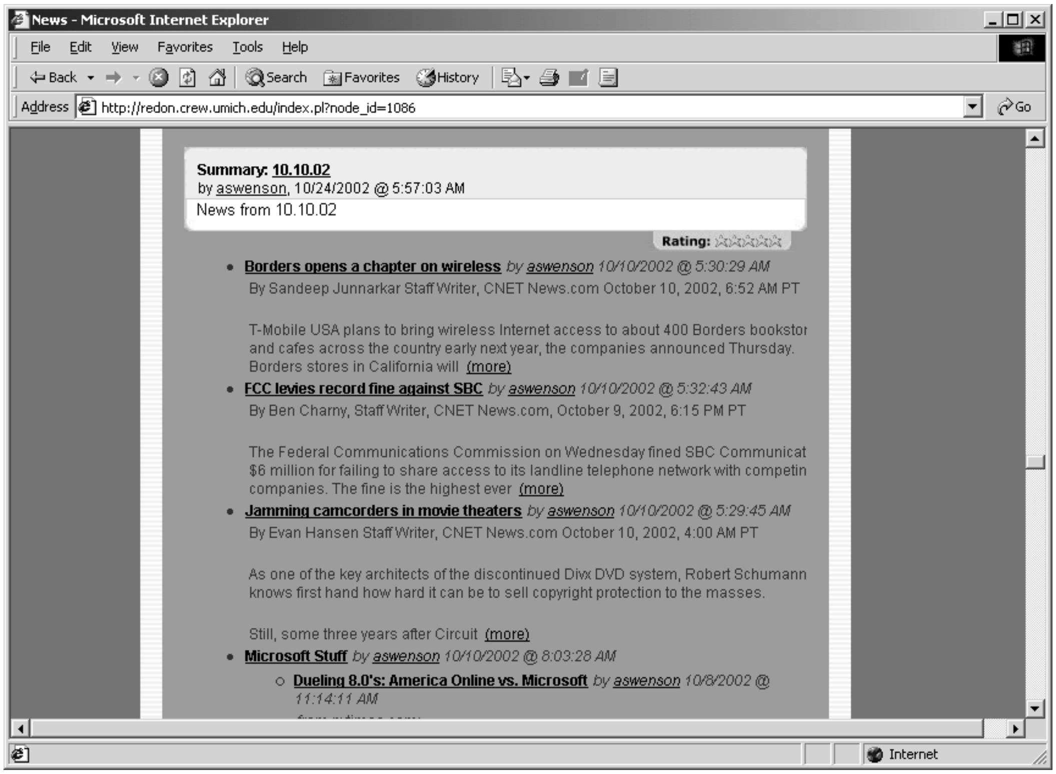
In addition to views of the information, Consolidate contains a query service used to find new relationships. The query service currently allows users to retrieve based on topic, date, keywords, and author. We believe that a major use will be retrieval by author. Many times one finds an unusually perspicacious or even offbeat author, and wishes to find other postings by the same author. In the future, we plan a “reduced keyword” query based on latent semantic indexing. In this query, both the message space and the query are mapped to an approximately 100 dimensional space; this can improve retrieval, especially for short messages.

Consolidate also contains a number of semi-autonomous agents. Some will be used to crunch visualizations of the messages. Editors must search for outliers, either to eliminate them from a consolidated site or to make them prominent because they have novel or offbeat ideas.

Implementation

Currently, both applications have been built. CyberForum consists of over 40,000 lines of Perl code, over and above the base Everything2 engine and our extensions. Our prototypes external agents are written in Java and are currently rather minor; the largest is several hundred lines of Java code. I-DIAG/Consolidate is constructed in Java and Jython, the Java implementation of the Python language. (Consolidate uses Jython as both an internal scripting language as well as a scripting language for user-created agents and user-modified views.) Consolidate currently consists of approximately 16,000 lines of Java code. Consolidate runs on any Java platform; CyberForum requires Debian Linux, a MySQL database, and an Apache Web server.

Only CyberForum is ready for full deployment. We are currently testing CyberForum in a limited field study


News - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back
Search
Favorites
History
Address
Go
Internet

http://redon.crew.umich.edu/index.pl?node_id=1086

Summary: 10.10.02
by aswenson, 10/24/2002 @ 5:57:03 AM

News from 10.10.02

Rating: ☆☆☆☆☆

- **Borders opens a chapter on wireless** *by aswenson 10/10/2002 @ 5:30:29 AM*
  By Sandeep Junnarkar Staff Writer, CNET News.com October 10, 2002, 6:52 AM PT

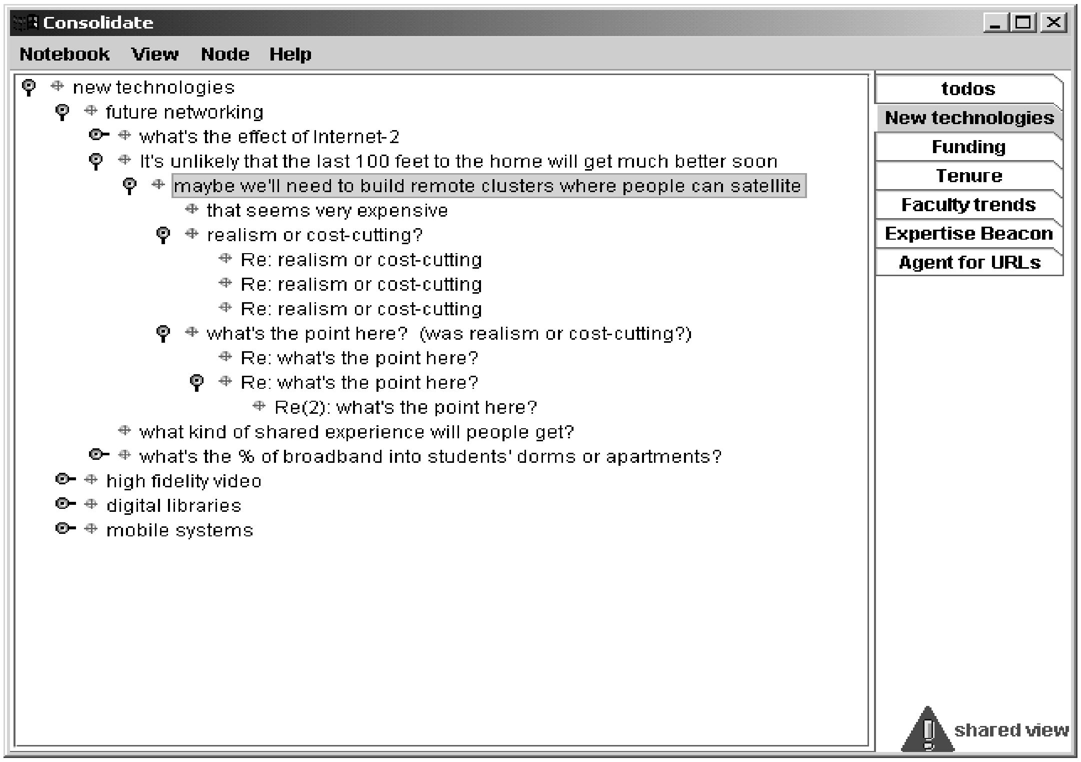
  T-Mobile USA plans to bring wireless Internet access to about 400 Borders bookstor
  and cafes across the country early next year, the companies announced Thursday.
  Borders stores in California will (more)
- **FCC levies record fine against SBC** *by aswenson 10/10/2002 @ 5:32:43 AM*
  By Ben Charny, Staff Writer, CNET News.com, October 9, 2002, 6:15 PM PT

  The Federal Communications Commission on Wednesday fined SBC Communicat
  $6 million for failing to share access to its landline telephone network with competin
  companies. The fine is the highest ever (more)
- **Jamming camcorders in movie theaters** *by aswenson 10/10/2002 @ 5:29:45 AM*
  By Evan Hansen Staff Writer, CNET News.com October 10, 2002, 4:00 AM PT

  As one of the key architects of the discontinued Divx DVD system, Robert Schumann
  knows first hand how hard it can be to sell copyright protection to the masses.

  Still, some three years after Circuit (more)
- **Microsoft Stuff** *by aswenson 10/10/2002 @ 8:03:28 AM*
  - **Dueling 8.0's: America Online vs. Microsoft** *by aswenson 10/8/2002 @ 11:14:11 AM*

Figure 3: CyberForum summary and messages



- new technologies
  - future networking
    - what's the effect of Internet-2
    - It's unlikely that the last 100 feet to the home will get much better soon
      - maybe we'll need to build remote clusters where people can satellite
        - that seems very expensive
        - realism or cost-cutting?
          - Re: realism or cost-cutting
          - Re: realism or cost-cutting
          - Re: realism or cost-cutting
        - what's the point here? (was realism or cost-cutting?)
          - Re: what's the point here?
          - Re: what's the point here?
            - Re(2): what's the point here?
    - what kind of shared experience will people get?
    - what's the % of broadband into students' dorms or apartments?
  - high fidelity video
  - digital libraries
  - mobile systems

Figure 4: A Consolidate viewsheet

consisting of two University of Michigan classes. We are planning a larger scale field test in the near future using both CyberForum and Consolidate.

Acknowledgements

This work has been funded, in part, by the National Science Foundation (IRI-9702904) and the University of Michigan/School of Information Alliance for Community Technology. Portions of I-DIAG Consolidate were originally funded by the MIT/Laboratory for Computer Science Project Oxygen Partnership. We would like to thank Dan Atkins, Cliff Lampe, and George Furnas for their insights on these issues.

References

Ackerman, Mark S., and David W. McDonald. 1996. Answer Garden 2: Merging Organizational Memory with Collective Help. *Proceedings of the ACM Conference on Computer-Supported Cooperative Work (CSCW'96)* : 97-105.

Ackerman, Mark S., and Leysia Palen. 1996. The Zephyr Help Instance: Promoting Ongoing Activity in a CSCW System. *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI'96)* : 268-275.

Conklin, Jeff, and Michael L. Begeman. 1988. gIBIS: A Hypertext Tool for Exploratory Policy Discussion. *Proceedings of the CSCW '88* : 140-152.

Dennis, Alan R., Jay E. Aronson, William G. Heninger, and Edward D. Walker. 1999. Structuring time and task in electronic brainstorming. 23 (1) : 95-108.

Hurwitz, Roger, and John C. Mallery. 1995. The Open Meeting: A Web-Based System for Conferencing and Collaboration. *Proceedings of the The Fourth International Conference on The World-Wide Web*

Jonassen, David H., Katherine Beissner, and Michael yacci. 1993. *Structural Knowledge: Techniques for Representing, Conveying, and Acquiring Structural Knowledge*. Hillsdale, NJ: Lawrence Erlbaum Associates.

MacLean, Allan, Richard Young, Victoria Bellotti, and Thomas Moran. 1990. Questions, Options, and Criteria: Elements of a Design Rationale for User Interfaces. EuroPARC/AMODEUS.

Moran, Thomas P., and John M. Carroll. 1996. *Design Rationale: Concepts, Techniques, and Use*. Lawrence Erlbaum.

Nunamaker, J. F., Alan R. Dennis, Joseph S. Valacich, Douglas Vogel, and Joey F. George. 1991. Electronic meeting systems. 34 (7) : 40-61.

Preece, Jenny. 2000. *Online Communities*. New York: Wiley.

Radev, Dragomir R., and Eduard Hovy. 1999. Intelligent text summarization. 3

Shipman, Frank, and Cathy Marshall. 1999. Formality Considered Harmful: Experiences, Emerging Themes, and Directions on the Use of Formal Representations in Interactive Systems. 8 (4) : 333-352.

Shipman, Frank, and Ray McCall. 1999. Supporting Incremental Formalization with the Hyper-Object Substrate. 17 (2) : 199-227.

Shipman, Frank M., III, and Raymond McCall. 1994. Supporting Knowledge-Base Evolution with Incremental Formalization. *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI'94)* : 285-291.

Sproull, Lee, and Sara Kiesler. 1991. *Connections: New Ways of Working in the Networked Organization*. Cambridge, MA: MIT Press.

Streitz, Norbert A., Jörg Geißler, Jörg M. Haake, and Jeroen Hol. 1994. DOLPHIN: integrated meeting support across local and remote desktop environments and LiveBoards. *Proceedings of the ACM conference on Computer supported cooperative work (CSCW'94)* : 345-358.

Webster's. 1986. *Webster's Ninth New Collegiate Dictionary*. Springfield, MA: Merriam-Webster.

Wenger, Etienne. 1998. *Communities of practice : learning, meaning, and identity*. New York: Cambridge University Press.