Supporting distributed scientific collaboration: Implications for designing the CiteSeer collaboratory

Umer Farooq[*], Craig H. Ganoe[*], John M. Carroll[*], and C. Lee Giles[+]
[*]*Computer Supported Collaboration and Learning Lab, Center for HCI*
[+]*The Intelligent Systems Research Lab*
College of Information Sciences and Technology
The Pennsylvania State University, University Park, PA 16802 USA
{ufarooq, cganoe, jcarroll, giles}@ist.psu.edu

Abstract

It is unclear if and how collaboratories have enhanced distributed scientific collaboration. Furthermore, little is known in the way of design strategies to support such collaboration. Based on a survey and follow-up interviews with CiteSeer users, we present four novel implications for designing the CiteSeer collaboratory. First, visualize query-based social networks to identify scholarly communities of interest. Second, provide online collaborative tool support for upstream stages of scientific collaboration. Third, support activity awareness for staying cognizant of online scientific activities. Fourth, use notification systems to convey scientific activity awareness.

1. Introduction

Scientific communities have traditionally formed around key intellectual resources such as collections of books, or special equipment such as cyclotrons [24]. In the past, one of the greatest obstacles to the formation and sustained vitality of scientific communities was the fact that members had to be co-located with their shared resources and with one another.

Today, face-to-face scientific collaboration is increasingly being augmented by online interactions. Collaboratories—laboratories for collaboration—enable large-scale scientific endeavors through Internet technologies. Through such environments, scientists can share key intellectual resources that allow colleagues located anywhere to access, view, manipulate, and have discussions about these artifacts [8, 16].

Although collaboratories have the potential to enhance distributed scientific collaboration, not much empirical evidence bears any mark of this. This can be partially attributed to the fact that most collaboratories have been built as one-off, handcrafted projects and have thus been accepted as the status quo [7, 20]. Furthermore, little is known in the way of design strategies to support distributed scientific collaboration. This is because only a few collaboratories have been evaluated from this angle (e.g., [19, 21], resulting in just a handful of basic design issues and heuristics related to general collaborative experiences in collaboratories (e.g., [7, 19]). In this paper, we present more specific design strategies geared toward scientific activities in distributed collaboration.

We report the first phase (requirements) of our research investigation to design a collaboratory around an existing, large-scale digital library of scientific literature in computing, namely CiteSeer (http://citeseer.ist.psu.edu). Based on a survey and follow-up interviews with CiteSeer users, we present four implications for design in order to support distributed scientific collaboration. These implications are novel in that they extend current literature in the domains of Human Computer Interaction (HCI) and Computer Supported Cooperative Work (CSCW).

2. Related work

The US, through its National Science Foundation (NSF), has been involved in collaboratory initiatives. A collaboratory is a "center without walls, in which the nation's researchers can perform their research without regard to geographical location—interacting with colleagues, accessing instrumentation, sharing data and computational resource, and accessing information in digital libraries" [26]. The challenges and opportunities in creating collaboratories and their interfaces relate directly to many aspects of HCI and CSCW. As a result of collaboratory development and HCI/CSCW research converging, a special issue of ACM Interactions was published in 1998, comprising four key articles that offered an in-depth look at collaboratories. An online list of collaboratories is also available [20].

Because the literature on collaboratories is so vast and scattered, it is appropriate to summarize the major

1530-1605/07 $20.00 © 2007 IEEE

findings from prior work. In 2002, Finholt [7] wrote a retrospective article in which he outlined a number of design issues related to collaboratory development. More recently in 2006, Olson and colleagues [19] attempted to propose a theory of remote collaboration based largely on their experience with collaboratory development. These two articles represent the state-of-the-art in designing collaboratories to support distributed scientific collaboration. We have codified the major findings from these sources in Table 1.

Table 1. Major findings from prior work.

Summary	*Source*
Tightly coupled tasks require co-location.	[18]
Collaboration "readiness" and "technology readiness" are essential factors for success.	[17]
Status misalignment can hamper communication between scientists.	[7]
Lack of common ground and trust can hinder collaboration.	[19]
Management, planning, and decision-making are critical processes to provide support for.	[19]
Allow flexibility to users for identifying new uses or functionality of tools.	[19]

3. Background of CiteSeer

Our study context is CiteSeer [9]: a search engine and digital library of literature in the computer and information science (CIS) disciplines that is a free resource providing access to the full-text of nearly 700,000 academic papers, and over 10 million citations. CiteSeer currently receives over half a million hits a day and is accessed by 150 countries and 200,000 unique machines monthly.

It is traditional practice in the CIS scientific community to make research documents available at the time they are first written through technical reports series managed by various laboratories. More recently, this practice has been transferred to the Web. CiteSeer actively and automatically harvests these documents and builds searchable and indexable collections, promoting creative scientific discovery and reuse. Even though search engines such as Google actively index CiteSeer, users come to CiteSeer for information such as citation counts and domain dependent citation links not provided by Google or Google Scholar.

4. Methods

Because CiteSeer has a large number of globally distributed users, we chose to administer an online survey. Broadly, we wanted to gain insight into the kinds of activities CiteSeer users would like to collaborate on and possible socio-technical issues during such collaboration.

4.1. Recruitment and participants

The survey was made available on CiteSeer's web site. Thus, this was an opportunity sample: participants were CiteSeer users willing to take the survey. No compensation was provided to survey respondents.

In this paper, we report results based on the administered survey during two weeks (November 17-30, 2005). 301 CiteSeer users responded to the survey.

4.2. Survey design

We report our results based on 23 survey questions organized into three broad sections. (1) *Professional interaction*: seven questions related to how users would like to collaborate with others and what issues they might face. (2) *CiteSeer use*: seven questions related to CiteSeer usage behavior. (3) *Background information*: nine questions related to demographics of CiteSeer users.

The questions were predominantly a mix of selection among pre-defined categories (e.g., age ranges) and ratings on 7-point Likert scales (e.g., engagement in a specific activity on a scale of "Never" to "Very often"); few free-text opportunities were provided (e.g., academic background). Based on pilot testing, the survey required 10-15 minutes to complete.

4.3. Data collection and analysis

Most survey questions solicited numerical responses. Analysis of this data was done using SPSS. Because we included multi-part questions in the survey, it was important to check the reliability of the scales. The scales on the multi-part questions had good internal consistency, with all Cronbach alpha coefficients reported above 0.7.

We wanted to probe user responses in more detail in order to complement the quantitative data. The second to last question asked for any type of qualitative feedback from participants (e.g., related to CiteSeer); 94 participants responded. The last survey question asked participants if they were willing to be interviewed via email.

We contacted 66 of these participants and got 22 responses. We asked four questions in the email interview: (1) Which criteria would you find most important for collaborating with CiteSeer users? (2) Which online collaborative activities would be most valuable to you? (3) Which activities would you like to stay most aware of? (4) What would be the best way for you to stay aware of these activities?

5. Survey results

Before reporting the results, we characterize the survey respondents with respect to their demographics and patterns of CiteSeer usage.

Of the responses we received, 42% were graduate students. Males (89%) outnumbered females. More than half the respondents (52%) were in the age range of 21-30 years.

42% of the respondents had a master's degree. The sample as a whole was relatively highly educated, with 32% having a doctorate degree. Because CiteSeer is a digital library primarily for the CIS disciplines, it was consistent that 79% of the respondents had at least a computer science background.

The survey respondents represented a relatively core group of CiteSeer users. Their mean (M) use of CiteSeer was 3.7 years (SD=1.7). Almost half (45%) had downloaded more than 100 papers from CiteSeer. 40% said they use CiteSeer once or twice per week.

We present the results under the following three themes: (1) *Potential collaborators*; (2) *Online collaborative activitie*s; and (3) *Awareness issues*.

5.1. Potential collaborators

We wanted to understand with whom CiteSeer users would collaborate online. Participants were asked to rate how often they would like to interact remotely with others on a scale of 1 (Never) to 7 (Very often) based on six items: (1) Who are looking for similar types of papers as I am; (2) Who read my papers; (3) Whose papers I read; (4) Who cite my papers; (5) Whom I cite in my papers; (6) Who cite similar papers as I do.

The six items were rated relatively high with all means above 4 (Sometimes): 4.76, 5.03, 5.10, 5.00, 4.97, and 4.65 respectively. Because the quantitative data is inconclusive, it is unclear which of the six criteria will be most useful to match potential collaborators. Qualitative data prioritizes some of these criteria.

For example, people are likely to collaborate with those who look for similar papers and read each other's papers. Reading similar papers is an indicator of people working in the same area, as one respondent suggests:

"Important criteria: users who are reading the same and similar papers as me. Since we are reading the same papers, we are working in the exact same sub-area."

It seems plausible that someone who looks for similar papers as another person also cites similar papers. In this case, potential collaborators can share common ideas that focus on the papers they look for or cite. One interview respondent expressed this view:

"[I want to] collaborate with CiteSeer users who are looking for similar papers as [me] and who cite similar papers as [I do]...the reason is I can save more time to find a good paper worth reading and can touch more ideas in my research area by collaboration."

A concern in matching people based on readings or citations is the use of personal, sensitive information. Surprisingly, no one indicated that using personal information would be an issue. On the contrary, one interview respondent suggested that people's web sites could be used to identify potential collaborators:

"[For] connecting users with common interests...focus on researchers' home pages, because almost everyone I have seen from academia gives a links page..."

One interview respondent provided an insight into how matching potential collaborators can also facilitate opportunistic collaboration outside of one's research area and expertise:

"[An] important aspect to collaboration is to facilitate 'serendipitous' interaction. As it is said, it's not what you don't know, it's what you don't know that you don't know. This is closely related to the discovery of cross domain knowledge and expertise."

The quantitative part of the survey did not probe users about the representation of social matching. As indicated by many survey respondents, social networks are appropriate for depicting meaningful social structures in CiteSeer:

"I think it would be great if I could get a CiteSeer page with a 'network' diagram...and 'related' strong links and more remote links clearly shown."

5.2. Online collaborative activities

We wanted to know what kinds of collaborative activities they would like supported. Participants were asked to rate how often they currently interact with others on a scale of 1 (Never) to 7 (Very often) based on four items: (1) Strengthen social connections; (2) Brainstorm new ideas; (3) Plan joint projects; (4) Write joint papers.

In general, respondents rated all items moderately high with all means above 4 (Sometimes): 4.28, 4.71, 4.32, and 4.06 respectively. Participants were also asked how difficult they would find these activities to achieve remotely on a scale of 1 (Very easy) to 7 (Very difficult). Responses indicate that CiteSeer users found these distributed collaborative activities to be on the difficult side of neutral (4), with respective means as 4.40, 4.36, 4.53, and 4.47.

One interpretation of these results is that CiteSeer users moderately engage in these types of collaborative activities. However, remote collaboration is perceived as somewhat difficult. Qualitative data elaborates on the kinds of online activities that CiteSeer users would like supported and gives reasons for not supporting other activities that they perceive as difficult.

Overwhelmingly, online discussions forums were the most popular type of distributed collaborative activity, as indicated by one of many respondents:

"I'd love to participate in forums or discussions about my field, to see what is going on, and what other people think."

Discussions can also be a valuable source for new ideas. The following interview respondent indicated the fact that discussions can enable brainstorming:

"[I would be interested in] brainstorming new ideas related to online discussions."

Given that CiteSeer users collaborate with others in collaborative planning and writing endeavors, these activities should be supported online. However, according to our interview respondents, they are not inclined to use such collaborative features. One interview respondent said:

"Writing new papers and planning projects don't seem like activities people would actually do through a science portal."

This respondent's view was corroborated by others who thought that current ways (e.g., email) of achieving such joint endeavors would suffice:

"I think the online discussions and brainstorming could be useful. For paper writing and project planning, I'd imagine that the team would be cohesive and we'd just use email or a wiki to coordinate."

Trust and privacy are obvious factors in hampering distributed collaboration. One respondent said:

"Collaboration is based on mutual trust, and it cannot be gained easily via an Internet site. Also, the question of privacy comes to my mind—one would not be willing to share his preliminary ideas to an unknown audience."

Establishing trust and privacy are exacerbated when potentially valuable ideas, which form the basis of scientific discovery, cannot be shared due to institutional constraints, or are shared and unethically misused. For example, legal issues can hinder distributed collaboration, as indicated by the following interview respondent:

"Some people will, no doubt, wish to be 'silent participants' [in online collaboration] due to legal intellectual property issues."

5.3. Awareness issues

We wanted to understand awareness issues in online collaboration and general CiteSeer use. Participants were asked to rate their level of agreement on how difficult they find it to stay aware of CiteSeer resources on a scale of 1 (Strongly disagree) to 7 (Strongly agree) based on four items: (1) Recent papers published in my area; (2) Who reads my papers; (3) New colleagues who are working in my area; (4) Who cites my papers.

Results suggest that staying aware was generally difficult as at least 50% of all respondents rated all items toward the agreement side of the scale. One-way within-subjects ANOVA was conducted with the awareness resources as the independent variable with four levels (the response items) and level of difficulty (rating from 1 to 7) as the dependent variable.

The Levene test was significant at 0.001, so the assumption of homogeneity of variance was violated. Therefore, both Brown-Forsythe and Welch F-ratios are reported. The ANOVA was significant, with $F(3, 594.44) = 22.68$ ($p<.0005$) and $F(3, 1057.04) = 22.08$ ($p<.0005$) respectively. We computed a contrast test between the first item (recent papers published in my area) and the other three items combined. Results indicate that the first item was rated significantly lower, with $F(1, 472.07) = 37.27$ ($p<.0005$). Thus, CiteSeer users find it less difficult to stay aware of recently published papers in their area, perhaps because this is done traditionally (through subscriptions to journals and conference attendance).

Although our quantitative questions only asked about the difficulty in staying aware, qualitative data suggests that awareness of CiteSeer resources and activities of CiteSeer users around those resources is important. An interview respondent said:

"[The most interesting awareness feature is] providing statistics on your own papers (readers, citations)."

Staying aware of new colleagues in one's research area is also important to keep abreast of potential collaborators, their activities, and their research focus. An interview respondent said:

"I'd like to know who has started a new discussion thread related to my area of interest, because I want to be aware what is going on outside my lab, and what other researchers are thinking or focusing on."

Qualitative data also suggests that mining historical activities in CiteSeer to provide influence patterns and impact assessment of intellectual resources can enrich awareness information. An interview respondent indicated the relevance of history for awareness and how it can also inform future impact of a discipline:

"It's always important to be aware of new research efforts starting up that are synergistic or disruptive relative to your own. You might consider online 'analytics' that give people some idea of where activity is centered and where it is going…It could tell you if, for example, interest in a discipline is 'dying down' or 'ramping up'."

In addition to historical information, supporting awareness of future activities is important to stay cognizant of current information. For instance, CiteSeer users want to be notified when a specific event has taken place, as indicated below:

"I would find it more important to know when a paper was entered into CiteSeer that cited one of my papers; that would be a strong signal that I might have interest in it."

An important facet of awareness is how it will be conveyed. Many interview respondents indicated the usefulness of Really Simple Syndication (RSS) feeds:

"[I] definitely [want] RSS: it isn't intrusive (I get information when I want), information can be easily [and] automatically processed, [and] I can get information in whatever way I want (as emails, in my aggregator, in my browser, ...)."

In addition to how awareness information can be conveyed, respondents indicated different types of information they would like to stay aware of. One respondent wanted to know about "hot topics" (implying popular topics) being discussed in forums. In another example, a respondent was interested in papers for a specified area of interest (e.g., using keywords) or those that cite his/her work:

"Features that would be useful are alerts when new articles are posted that either contain keywords or cite work I am interested in to keep abreast of what's new in my field."

Even though there are traditional ways of staying aware of new papers, using features to refine such awareness (e.g., through keywords) seems desirable.

6. Implications for design

Several of our results suggest specific strategies to support distributed scientific collaboration. The four implications for design are the following. (1) *Visualize query-based social networks to identify scholarly communities of interest*. (2) *Provide online collaborative tool support for upstream stages of scientific collaboration*. (3) *Support activity awareness to stay cognizant of online, asynchronous, and long-term scientific activities*. (4) *Use notification systems to convey scientific activity awareness peripheral to users' primary task*.

The implications are motivated by design rationale based on survey results and related HCI/CSCW literature. Design envisionment scenarios, conceptual schemas, and prototype screenshots are used to illustrate the implications for design.

6.1. Visualize query-based social networks

In regard to matching potential collaborators, survey results support existing claims. Literature from social psychology asserts that people are attracted to "similar others" [23, p. 416], with similarity in interests being one facet of this. In CiteSeer, identifying users with similar interests can be based on multiple criteria, such as mutual reading of papers, citations, and similar search behavior. Similar search behavior seems to be a feasible candidate among these choices for at least three reasons.

First, CiteSeer can easily keep track of users' search behavior by storing and mining a history of user queries. CiteSeer queries—typically, noun phrases such as "user-centered design"—essentially filter the space of available resources into specialized views. These views can be thought of as research investigations, research areas, or even sub-disciplines. Many queries are in effect reused in the sense that someone else entered that query, or one like it, before. Comparing these queries with similarity measures can provide social matching heuristics for users.

Second, search queries are universal. For example, social matching based on citations may not apply to all users as everyone would not have a critical mass of cited papers (e.g., graduate students).

Third, queries accurately convey first-hand information about a user's interests. Queries that cumulate over time related to the same topic can indicate a strong interest in that topic. Of course, two users submitting similar queries do not necessarily want to collaborate, but the chance that collaboration would be attractive at some level is more likely than individuals with totally different interests.

Scholarly communities and sub-communities can form around queries, just as they have traditionally formed around shared resources. Providing a virtual place for scientists with common query interests to share perspectives, related and updated information and links, and so forth would enrich these queries for everyone, and help scholarship and scholarly communities of interest or practice to form and develop [25].

These scholarly communities could be codified through social network analysis where shared queries are the primary basis for links among persons in the network. Query-based social networks would connect persons more or less directly, depending on how many queries they shared, and how they were connected to others in the network. We might expect interesting community phenomena to emerge from such networks. For example, the network could foster scientific collaboration, not just between members within a particular scientific group, but also between *weak ties* [10], scholars principally belonging to different groups who are connected through others. This can help CiteSeer users to identify new colleagues and potential collaborators more easily.

Social structures can also be used to discover and reinforce cross-community *bridges*. Bridges are, at the most basic level, members of two or more distinct community organizations [14]. In a scientific community, bridges are researchers who are part of two or more research communities (e.g., HCI and IS: Information Systems). Through query-based social networks, scientists can opportunistically explore nodes and edges beyond their immediate task goals, and learn about bridges and their expertise that complement their own research area. Scientists who expand the edges of their communities in this way become more aware of activities that might influence their own work. This perspective aligns well with our survey results that indicated the advantages of serendipitous collaboration.

An issue in social matching systems is the use of personal information. Personal information is critical for matching people. Terveen and McDonald [23] claim that social matching systems need to use—and users will be willing to supply—relatively personal sensitive information to effectively match people. It is worthwhile to note that while Terveen and McDonald meant personal sensitive information to imply age, music taste, hobbies, and so forth, we are construing such information for scientific communities as user queries. We anticipate few problems in getting scientists to allow use of their queries (anonymous to other users) for system-level social matching, given that evidence suggests that people will be willing to share more personal sensitive information, as per Terveen and McDonald's claim. Survey results mildly indicate that users would be willing to provide such information, such as their personal web sites.

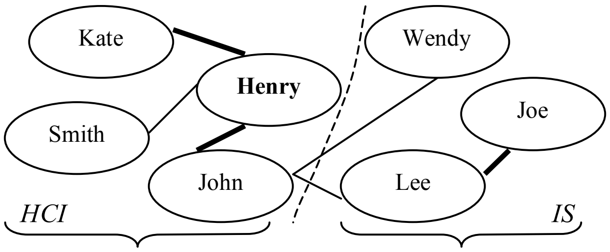


Figure 1. Conceptual schema of a social network.

Example. Consider the conceptual schema of a possible query-based social network in Figure 1. Henry is strongly connected to Kate and John based on many similar queries they share (connections shown in bold) and to Smith based on few similar queries (non-bold connections) in the area of HCI. John happens to be connected to Lee and Wendy based on his similar interest in the area of IS. Because John is connected with people in both areas of HCI and IS, he is a bridge. Henry is surprised to know that his research questions are also addressed in IS literature. Henry can leverage John as a bridge to see how his research in HCI fits into IS paradigms. Henry is excited to contact Lee or Wendy who happen to be his weak ties through John. He feels that cross-community collaboration can enrich his theoretical work.

6.2. Support upstream stages of collaboration

Survey results suggest that CiteSeer users would welcome opportunities that support open-ended and idea-generation activities. Contrary to focused activities, we characterize such opportunities as upstream stages of scientific collaboration. This refers to early, divergent stages of scientific discovery in contrast to final, convergent stages.

While asynchronous discussion forums are ideal for open-ended and divergent technical discussions, they are often not flexible or interactive enough to support finer-grain collaborations like joint authoring [15]. Thus, it was our intention for CiteSeer users to be able to create collaborative spaces for more synchronous, sustained, and convergent collaborative interactions for developing intellectual products such as research papers and proposals. However, survey results strongly suggested against such focused tool support.

In addition to issues such as trust and privacy, results indicated that users did not want support for such focused collaborative activities because they already had existing ways of engaging in such endeavors. Some respondents suggested that face-to-face interactions and email are sufficient to achieve focused activities. Hollan and Stornetta [12] assert that face-to-face interactions cannot be replaced by any other collaboration channel, and therefore, the goal of developing tools for distributed interaction should be to identify needs that are not met in physical proximity. Olson and Olson [18] also suggest that co-location is still essential for some collaboration, especially for tightly coupled and focused activities that demand frequent interaction and feedback among participants.

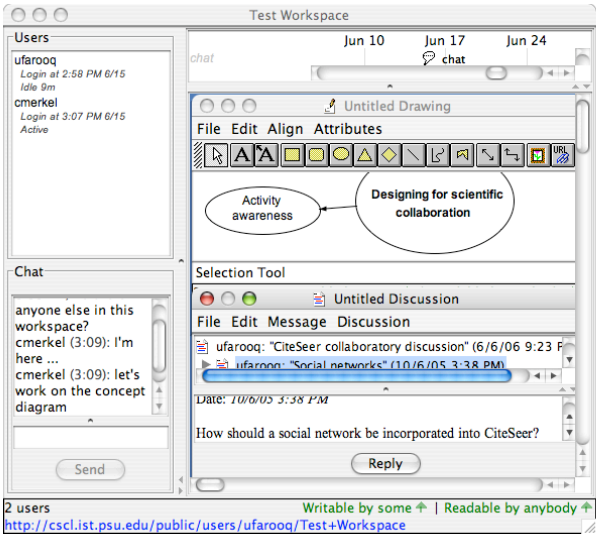
The design rationale for supporting upstream stages of scientific collaboration stems from at least two reasons. First, lack of common ground, trust, and intellectual ownership should be less important issues at the preparatory rather than concluding stages of scientific collaboration. This is because the goal during upstream stages of collaboration is to generate, share, and leverage ideas with a communal orientation. The benefits of collectively engaging in such collaboration are likely to outweigh its costs.

Second, supporting upstream stages of scientific collaboration represents a segue from just search and retrieval tasks of CiteSeer's resources to interacting minimally with other users. This is consistent with the existing finding that technology readiness is required for successful collaboration [17]. Attempting to leapfrog steps by providing sophisticated applications (e.g., collaborative writing tools) rather than progressive interventions can produce frustration and resistance on part of the users.

Survey results provided examples of tools that could support upstream stages of scientific collaboration. Discussion-oriented tools were a popular demand. CiteSeer currently can directly present the influence network for a resource (e.g., a listing of papers that cite Grudin's paper "Groupware and social dynamics: Eight challenges for developers"), but it does not provide a textual exegesis synthesizing and interpreting that network of citations (e.g., discussions on the ideas in Grudin's paper, their influence on particular researchers, etc). Such an exegesis could be the social construction of a scientific community. Providing an explicit medium to codify such discussions can enrich the specific resources for everyone who accesses them, and more generally can help scholarship and scientific communities develop.

In addition to discussion tools, collaborative brainstorming tools such as concept maps and white-boards are likely to support scientific discovery. It has been shown that brainstorming can increase the ability to share and generate creative ideas [22].

Example. Currently, we are prototyping a workspace that supports upstream stages of scientific collaboration within CiteSeer. Figure 2 shows a screenshot of this prototype. The idea is that CiteSeer users can engage in synchronous and asynchronous brainstorming activities, such as through collaborative concept maps and threaded discussions. The timeline on the top maintains version histories of collaborative activities (example of chat session on the left).



anyone else in this workspace?
cmerkel (3:09): I'm here ...
cmerkel (3:09): let's work on the concept diagram

ufarooq: "CiteSeer collaboratory discussion" (6/6/06 9:23 F
ufarooq: "Social networks" (10/6/05 3:38 PM)

Date: 10/6/05 3:38 PM

How should a social network be incorporated into CiteSeer?

http://cscl.ist.psu.edu/public/users/ufarooq/Test+Workspace

Figure 2. Collaborative workspace prototype.

6.3. Support activity awareness

CSCW literature has highlighted the importance of awareness for successful collaboration [5]. For example, it is critical to know who else is present—social awareness [6]—and what others are doing—workspace awareness [11]—in a shared workspace.

Survey results suggest that supporting awareness for CiteSeer users is an opportunity. The type of desired awareness features expressed by survey participants cannot be adequately supported by traditional types of CSCW awareness mechanisms. This is because awareness in CSCW has focused more on supporting synchronous mediums of interactions over brief periods of time: awareness of who is participating in an ongoing activity, awareness of what each person is currently doing in that activity context, and awareness of how the team as a whole is performing [4]. Asynchronous and long-term awareness phenomena have been investigated somewhat less. Furthermore, investigating awareness specifically for scientific collaboration has not been explored before.

Survey results suggest that for CiteSeer, most activities are asynchronous and long-term. For instance, users use CiteSeer intermittently over the course of months and years, depending on when they need to access intellectual resources. This implies that although staying aware of what is going on at the present time is important, awareness of historical and future activities is key to successful collaboration.

Activity awareness [3] has sought out to provide such activity-based information. Activity awareness is awareness of project work that supports group performance in complex tasks over long-term endeavors directed at major goals. Activity awareness allows reflection of one's work, review of prior session histories, and analysis of future collaborative endeavors.

The design rationale for specifically using activity awareness is grounded in activity theory (see [3] for details). An activity-centered perspective emphasizes complex socially and culturally embedded endeavors that are organized in dynamic hierarchies. For example, scientific activities involve convergent and divergent thinking, development of professional relationships with peers, collaboration with others that unfolds over time, dissemination of work in the broader scientific community, and so forth.

The argument is that real world collaborators, such as scientists, need to be aware of one another's activity, understood in this broad sense. Carroll et al. [3] described a framework for activity awareness that takes the perspective of individuals operating within *communities of practice* (such as a scientific community) that emerged and are sustained through the construction of *common ground*, exchange of *social capital*, and the processes of *human development*. Such a framework is highly appropriate for supporting scientific activity awareness during distributed collaboration. This is because scientists have personal goals for contribution and reputation (human development), as they collaborate with peers (social capital) based on mutual trust and knowledge (common ground), operating in a globally distributed research environment (community of practice) [4].

Example. Consider the following scenario:

To assess the impact of her research, Diane signs up to be alerted when any of her papers in CiteSeer are cited. She also subscribes to a service for notifying her when new papers in her research area are available. While editing her paper, Diane receives a notification that Larry Somers has just published an article in her flagship journal. She shares this article with her graduate students to discuss how their proposed experiment can build on the article's empirical results.

In this scenario, scientific activity awareness is supported through a subscription service that computes influence patterns of papers based on citations. Scientific activity awareness was also used to keep track of latest research in a community of practice (community of scholars in one's research area). Here, Diane's immediate research is affected by the publication of a recent journal article that generates social capital in her research group. Activity awareness allows one to keep abreast of such online, asynchronous scientific activities over time.

6.4. Use notification systems

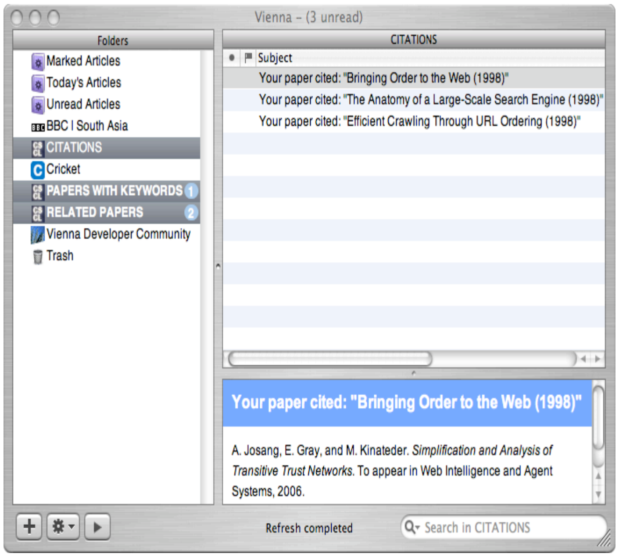
Part of the challenge in supporting computer-supported awareness is knowing *how* to convey it effectively. Survey results suggest that alerting services like RSS are preferred. We refer to such awareness mechanisms as notification systems.

Notification systems appear to be a reasonable mechanism to convey scientific activity awareness. Notification systems are typically lightweight, event-triggered displays of information peripheral to a person's current task-oriented concern, for example, system status updates, email alerts, stock tickers, and chat messaging. Notification systems have been used before to support collaborative activity awareness [2].

The design rationale for using notification systems to convey scientific activity awareness is based on at least two reasons. First, awareness of scientific activities is not the primary task of the user but peripheral to it. For example, in the activity awareness scenario (section 6.3), Diane wants to be alerted of status updates related to her citations or new papers; seldom will these be her primary activities. Because awareness of scientific activities is secondary to a user's primary task, it needs to be conveyed in a lightweight, non-intrusive way, yet be effective enough to capture the user's attention and cause some response. Notification systems fit exactly this profile.

Second, survey results indicated that flexibility is required in configuring not only *how* awareness information should be conveyed but also *what* should be conveyed. For example, some CiteSeer users would be interested in citations to their papers, others in when new papers are available, yet some would want to know when a specific discussion thread has been posted. Notification systems provide such flexible configurability so users get the right kind of information in the ways that they want it.

Recently, we have been exploring the use of object-based RSS feeds, as opposed to traditional news-based feeds, as notification systems to convey activity awareness in collaborative settings [13]. RSS feeds seem appropriate for conveying scientific activity awareness because of their configurable, non-intrusive, and lightweight nature. Also, many survey respondents provided strong support for RSS feeds in CiteSeer as indicated by our results.



Your paper cited: "Bringing Order to the Web (1998)"

Your paper cited: "The Anatomy of a Large-Scale Search Engine (1998)"

Your paper cited: "Efficient Crawling Through URL Ordering (1998)"

Your paper cited: "Bringing Order to the Web (1998)"

A. Josang, E. Gray, and M. Kinateder. *Simplification and Analysis of Transitive Trust Networks.* To appear in Web Intelligence and Agent Systems, 2006.

Refresh completed

Figure 3. RSS simulations in Vienna client.

Example. We are currently implementing a simulated prototype (Figure 3) for supporting scientific activity awareness through notifications systems based on RSS. The prototype uses a real newsreader client (Vienna).

The three types of simulated feeds being evaluated are highlighted in the figure: (1) *Citations*: citations to one's papers in CiteSeer; (2) *Papers with keywords*: CiteSeer papers related to one's specified keywords; and (3) *Related papers*: related papers to one's papers in CiteSeer. The prototype is being evaluated to gauge how RSS feeds can better support activity awareness.

7. Discussion and future work

Our findings emerged in context of a scientific community that exists around the intellectual resources of a digital library. Our implications for design can be applied to similar infrastructures as CiteSeer. For example, one could certainly imagine enhancing the ACM Digital Library (http://www.acm.org/dl/) as a collaboratory to support collaboration within the broader computer science scientific community. Our findings would certainly be within reasonable scope for such an undertaking. For example, using notification systems to support scientific activity awareness of latest research trends in the artificial intelligence sub-community is a likely scenario.

We do want to acknowledge two caveats regarding the survey sample. First, the opportunity sample may not be representative of the CiteSeer scientific community. We reconcile this shortcoming with the fact that self-selection was the only realistic sampling procedure available to us.

Second, we have not yet analyzed the data according to demographic factors. For example, scientists at different stages of their careers may have different needs for making new contacts and engaging in collaboration [1]. Stratifying and analyzing data by professional status, gender, educational background, and geographical location can enrich our interpretation. We plan to take demographic factors into account in future CiteSeer studies.

As immediate future work, we are enhancing CiteSeer's infrastructure to support some of the design strategies presented in this paper. From prior work, we have identified BRIDGE (Basic Resources for Integrated Distributed Group Environments; http://bridgetools.sourceforge.net) as a compatible environment to integrate with CiteSeer. BRIDGE already supports asynchronous, collaborative activities such as brainstorming, white-boarding, and concept mapping (some features were illustrated in Figure 2), and provides activity awareness through notification systems in context of collaborative work [2]. By gearing the BRIDGE functionality toward scientific collaboration, we plan to iteratively prototype and formatively evaluate the CiteSeer collaboratory.

8. Conclusion

Finholt [7] rightly points out that collaboratories represent an important convergence of computing technology with scientific practice. However, to qualitatively advance scientific practice, the design space requires novel and specific insights from more collaboratory case studies. Previous findings (Table 1) certainly inform us of factors that hamper successful distributed collaboration (e.g., lack of common ground, collaboration readiness, etc.) but they seldom specify design strategies to counter these issues. Our implications for design, emerging from the requirements phase of the proposed CiteSeer collaboratory, extend current findings.

The empirical results from the administered survey and follow-up interviews specifically raise issues of community building and collaboration for CiteSeer users. We found that users are inclined to interact with potential collaborators based on various criteria. Making such criteria visible, such as users having similar research interests through query-based social networks, can facilitate more meaningful collaboration.

We also reported that supporting collaborative activities in the early, divergent (upstream) stages of scientific discovery are a first approximation to enable collaboration currently between CiteSeer users. This can workaround issues such as trust and privacy for the time being until users become progressively motivated and confident in using collaborative tools that support more focused activities.

Finally, we found that users perceive CiteSeer as a resource for keeping aware of the vector of activities occurring in their field and others. Activity awareness through notification systems is a promising candidate for keeping track of long-term changes to intellectual resources and shared activities around those resources.

9. Acknowledgments

Our work is partially supported by US National Science Foundation (NSF) grant CRI-0454052. We thank Isaac Council and Levent Bolelli for setting up the survey on CiteSeer's web site. The first author would like to express gratitude to Thomas Kannampallil, Joey Lee, Joel Galbraith, Kyu Yon Lim, and Dana Mitra for providing much needed feedback on multiple revisions of this paper, and Sue Kase and once again Thomas for help in analyzing the data.

10. References

[1] A. Bruckman and C. Jenson, "The mystery of the death of MediaMoo: Seven years of evolution of an online

community", in K.A. Renninger & W. Shumar (Eds), *Building virtual communities: Learning and change in cyberspace*, Cambridge University Press, 2002, pp. 21-33.

[2] J.M. Carroll, D.C. Neale, P.L. Isenhour, M.B. Rosson, and S.D. McCrickard, "Notification and awareness: synchronizing task-oriented collaborative activity", *International Journal of Human-Computer Studies*, 58, 2003, pp. 605-632.

[3] J.M. Carroll, M.B. Rosson, G. Convertino, and C.H. Ganoe, "Awareness and teamwork in computer-supported collaboration", *Interacting with Computers*, 18 (1), 2006, pp. 21-46.

[4] J.M. Carroll, M.B. Rosson, U. Farooq, G. Convertino, C.B. Merkel, W.A. Schafer, C.H. Ganoe, and L. Xiao, "Beyond being aware", *Human Computer Interaction Consortium* (Fraser, Colorado, February 1-5, 2006).

[5] P. Dourish and V. Bellotti, "Awareness and coordination in shared workspaces", *Proceedings of the Conference on Computer Supported Cooperative Work*, (Toronto, Canada, October 31-November 4, 1992), ACM Press, New York, NY, 1992, pp. 107-113.

[6] T. Erickson, D.N. Smith, W.A. Kellogg, M.R. Laff, J.T. Richards, and E. Bradner, "Socially translucent systems: social proxies, persistent conversation, and the design of Babble", *Proceedings of the Conference on Human Factors in Computing Systems* (Pittsburg, PA, May 15-20, 1999). ACM Press, New York, NY, 1999, pp. 72-79.

[7] T.A. Finholt, "Collaboratories", in B. Cronin (Ed), *Annual Review of Information Science and Technology*, 36, ASIST, Washington DC, 2002, 73-107.

[8] T.A. Finholt and G.M. Olson, "From laboratories to collaboratories: A new organizational form for scientific collaboration", *Psychological Science*, 8 (1), 1997, pp. 28-35.

[9] C.L. Giles, K. Bollacker, and S. Lawrence, "CiteSeer: An Automatic Citation Indexing System", *Proceedings of the Conference on Digital Libraries* (Pittsburgh, PA, June 23-26, 1998), ACM Press, New York, NY, 1998, pp. 89-98.

[10] M. Granovetter, "The Strength of Weak Ties", *American Journal of Sociology*, 78 (6), 1973, pp. 1360-1380.

[11] C. Gutwin and S. Greenberg, "Workspace awareness for groupware", *Proceedings of the Conference on Human Factors in Computing Systems* (Vancouver, Canada, April 13-18, 1996), ACM Press, New York, NY, 1996, pp. 208-209.

[12] J. Hollan and S. Stornetta, "Beyond Being There", *Proceedings of the CHI Conference on Human Factors in Computing Systems* (Monterey, CA, May 3-7, 1992), ACM Press, New York, NY, 1992, pp. 119-125.

[13] K. Hylton, M.B. Rosson, J.M. Carroll, and C.H. Ganoe, "When news is more than what makes headlines", *ACM Crossroads*, 12 (2), 2005.

[14] A. Kavanaugh, D.D. Reese, J.M. Carroll, and M.B. Rosson, "Weak ties in networked communities", *The Information Society*, 21 (2), 2005, pp. 119-131.

[15] S.U. Kelly, C. Sung, and S. Farnham, "Designing for improved social responsibility, user participation and content in on-line communities", *Proceedings of Conference Human Factors and Computing Systems* (Minneapolis, MN, April 20-25, 2002), ACM Press, New York, NY, 2002, pp. 391-398.

[16] R.T. Kouzes, J.D. Myers, and W.A. Wulf, "Collaboratories: Doing science on the Internet", *IEEE Computer*, 29 (8), 1996, pp. 40-46.

[17] G.M. Olson, T.A. Finholt, and S.D. Teasley, "Behavioral aspects of collaboratories", in S.H. Koslow & M.F. Huerta (Eds), *Electronic collaboration in science*, Lawrence Erlbaum Associates, Mahwah, NJ, 2000, pp. 1-14.

[18] G.M. Olson and J.S. Olson, "Distance matters", *Human Computer Interaction*, 15, 2001, pp. 139-179.

[19] J.D. Olson, E. Hofer, D. Cooney, A. Zimmerman, N. Bos, and G.M. Olson, "Bridging distance in collaboration: Towards a theory of remote collaboration", *Human Computer Interaction Consortium* (Fraser, Colorado, February 1-5, 2006).

[20] "Science of Collaboratories", http://www.scienceofcollaboratories.org (accessed March 16, 2006).

[21] D.H. Sonnenwald, M.C. Whitton, and K.L. Maglaughlin, "Evaluating a Scientific Collaboratory: Results of a Controlled Experiment", *ACM Transactions on Computer Human Interaction*, 10 (2), 2003, pp. 150-176.

[22] R.I. Sutton and A. Hargadon, "Brainstorming groups in context: Effectiveness in a product design firm", *Administrative Science Quarterly*, 41, 1996, pp. 685-718.

[23] L. Terveen and D.W. McDonald, "Social matching: A framework and research agenda", *ACM Transactions on Computer Human Interaction*, 12 (3), 2005, pp. 401-434.

[24] Wellman, B. (Eds), *Networks in the global village: Life in contemporary communities*, Westview Press, Boulder, CO, 1999.

[25] Wenger, E., R. McDermott, and W. Snyder, *Cultivating communities of practice: A guide to managing knowledge*, Harvard Business School Press, Boston, MA, 2002.

[26] W.A. Wulf, "The Collaboratory Opportunity", *Science*, 261, 1993, pp. 854-855.