Available online at www.sciencedirect.com

ScienceDirect

Information Processing and Management 44 (2008) 596–612

INFORMATION PROCESSING & MANAGEMENT

www.elsevier.com/locate/infoproman

Design and evaluation of awareness mechanisms in CiteSeer

Umer Farooq *, Craig H. Ganoe, John M. Carroll, Isaac G. Councill, C. Lee Giles

College of Information Sciences and Technology, The Pennsylvania State University, University Park, PA 16802, USA

Received 25 October 2006; received in revised form 8 May 2007; accepted 18 May 2007
Available online 16 July 2007

Abstract

Awareness has been extensively studied in human computer interaction (HCI) and computer supported cooperative work (CSCW). The success of many collaborative systems hinges on effectively supporting awareness of different collaborators, their actions, and the process of creating shared work products. As digital libraries are increasingly becoming more than just repositories for information search and retrieval – essentially fostering collaboration among its community of users – awareness remains an unexplored research area in this domain. We are investigating awareness mechanisms in CiteSeer, a scholarly digital library for the computer and information science domain. CiteSeer users can be notified of new publication events (e.g., publication of a paper that cites one of their papers) using feeds as notification systems. We present three cumulative user studies – requirements elicitation, prototype evaluation, and naturalistic study – in the context of supporting CiteSeer feeds. Our results indicate that users prefer feeds that place target items in query-relevant contexts, and that preferred context varies with type of publication event. We found that users integrated feeds as part of their broader, everyday activities and used them as planning tools to collaborate with others.

Keywords: Computer-supported awareness; Notification systems; Communities of practice; Scholarly digital libraries

1. Introduction

Scholars have always organized themselves into communities, within which they collaborate, debate, and otherwise interact. Today, the number and variety of communities, forums, and channels for such interactions are vast. Indeed, the recent growth of interdisciplinary scholarship – particularly in the sciences – has made contemporary scholarly communities more open, but also less identifiable. The network of people and resources relevant to one's current scholarly project is quite dynamic; through the course of a career, or even of a project, one may move through a series of foci. Thus, it is now typical for scholars to wonder what communities are addressing issues of interest to them, and even what communities they are influencing.

A concrete way to think about this challenge is that contemporary scholars need to be *aware* of a far wider range of colleagues and research topics. But clearly the solution is not to thoroughly read or even manually browse ever-widening swaths of research. That is not humanly possible. Fortunately, these new needs are

[*] Corresponding author.
E-mail address: ufarooq@ist.psu.edu (U. Farooq).

0306-4573/$ - see front matter © 2007 Elsevier Ltd. All rights reserved.
doi:10.1016/j.ipm.2007.05.009

concurrent with powerful new infrastructures for research in the form of digital libraries, collaboratories, and the Internet homepages of researchers throughout the world. The answer to the challenge of how scholars can become aware of relevant colleagues and resources, then, can be addressed, at least in part, by investigating new sorts of digital tools to apprise them of events on the World Wide Web. For example, a paper has just been published that cites one's prior paper or matches one's keyword profile.

This approach takes the concept of "scholarly community" quite literally, and indeed extends it to encompass scholarly activity throughout the Web. This re-scoping of the challenge of scholarly awareness will not come as a surprise to contemporary scholars, but it entails substantial scaling of concepts and techniques. For example, awareness has been studied in the context of human computer interaction (HCI) and computer supported cooperative work (CSCW) for 15 years, but it has typically been conceived of with respect to relatively immediate and short-term events and processes in a relatively closed system supporting relatively small groups (e.g., Dourish & Bellotti, 1992). Extending the concept of awareness to digital libraries and to longer-term, perhaps even open-ended projects is only beginning (e.g., Hansen & Järvelin, 2005).

In this paper, we are interested in investigating the feasibility and effectiveness of feed-based notification systems as part of users' broader activities to stay aware of new publication events. Our study context is CiteSeer (http://citeseer.ist.psu.edu), a scholarly digital library of research literature in the computer and information science disciplines. To stay aware of new publication events (e.g., publication of a paper that cites your paper), *notification systems* (McCrickard, Chewar, Somervell, & Ndiwalana, 2003) can be used as awareness mechanisms to deliver relevant information to users in an efficient and effective manner. Currently, such functionality in digital libraries is primarily limited to alert services through email notifications (e.g., Google alerts; http://www.google.com/alerts). To this end, we present three cumulative user studies – requirements elicitation, prototype evaluation, and naturalistic study – in the context of supporting notification systems as awareness mechanisms in CiteSeer.

The first study surveyed CiteSeer users to elicit requirements for enhancing the scholarly digital library into a collaboratory; indeed, prior literature has placed emphasis on supporting collaboration in digital libraries (Blandford & Gow, 2006; Hansen & Järvelin, 2005; Lagoze, Krafft, Payette, & Jesuroga, 2005). Among many novel design requirements and implications that were articulated based on this study (reported fully in Farooq, Ganoe, Carroll, & Giles, 2007), one of the enhancements that users indicated was to support awareness of new publication events in CiteSeer through *RSS feeds* as notification systems.

RSS (really simple syndication or rich site summary) is an XML-based format that allows users to subscribe to updates on their favorite websites. For example, a news source like CNN can use RSS to broadcast news feed items that typically contain a title, a textual summary, timestamp, and a link to the news article. Using a feed reader, users can periodically get this information from the news source, thus receiving updates about new or changed items. In two subsequent studies, we investigated the design and evaluation of RSS feeds in CiteSeer.

In our second study, we presented RSS mock-up prototypes with different content designs to CiteSeer users for assessing their preferences for various types of publication event feeds. Based on these preferences, we conducted a third study with six teams whose members were asked to collaborate with each other in a distributed setting over several days. Group members worked together on a realistic, shared task using a collaborative workspace with the support of RSS feeds for CiteSeer. This naturalistic study allowed us to understand user behavior toward the CiteSeer RSS feeds in the context of their everyday activities. The balance of this paper motivates our research investigation and describes these three studies in detail.

2. Motivation and related work

From an HCI and CSCW perspective, the literature on digital libraries has been primarily techno-centric, taking user experience little into account (Blandford & Gow, 2006; Dillon, 2002). Behavioral studies of digital libraries have focused largely on individual users, following traditions of information-seeking research; only a limited number of studies (e.g., see studies in Bishop, Van House, & Buttonfield, 2003) have looked at the collaborative aspects of digital libraries (Borgman, 2006).

Recently, in the digital library domain, there has been an increasing focus on user practices and their social interactions with regard to digital library usage (Adams, Blandford, Budd, & Bailey, 2005). While collaboration is understood to be an important aspect of digital libraries, there is actually little *empirical* knowledge about collaboration and issues in collaboration within the realm of information search and retrieval processes (Hansen & Järvelin, 2005). For example, Renda and Straccia (2005) envision that digital libraries can indeed be considered as collaborative meeting and common working places where users become aware of each other, open communication channels, and exchange information and knowledge with each other or with experts. But their contribution, which is important in its own regard, is purely technical in nature.

Considering digital libraries as communities of practice and mediums for collaborative endeavors (Borgman, 1999), one of the major research issues to consider is awareness, given that it is a critical requirement for successful collaborative systems. Awareness has taken up many meanings and interpretations, and highly depends on the context for which it is being used (Schmidt, 2002). In a CSCW context, awareness tends to imply practices through which collaborators know the social context of work, such as what their colleagues are doing in a shared workspace (*workspace awareness*: Gutwin & Greenberg, 1996).

In the context of digital libraries, research on awareness has been scarce. Adams et al. (2005) have investigated *organizational awareness*, referring to awareness of community activities, events, and resources across an organization. They discuss the design and evaluation of a screen saver application as an awareness communication medium in the clinical domain.

Hansen and Järvelin (2005) describe awareness more formally, classifying it as awareness of *people*, *activities*, and *objects*. Awareness of people refers to knowing about one's colleagues. Awareness of activities refers to sharing the same need for information such as search strategies. Awareness of objects refers to accessing different types of resources such as sharing retrieved objects.

The work most closely related to ours is by Collins, Mane, Martinez, Hussell, and Luce (2005). They describe the design of ScienceSifter, a tool that enables researchers and scientists to create and customize information feeds. A unique feature in ScienceSifter is the ability of users to select multiple feeds, aggregate them into one feed, and then use a set of keywords to filter the feed. In this way, ScienceSifter has the potential to facilitate efficient information sharing, although no users studies were reported to validate the effectiveness of the tool. In our work, we are actually trying to understand how users would tailor content of the feeds to their preferences, rather than just customizing different types of feeds, and how feeds can be integrated in the context of users' broader activities.

In the domain of health informatics, researchers have explored the design and evaluation of alerting services for digital libraries. Buchanan and Hinze (2005) and Hinze, Buchanan, Jung, and Adam (2006) identify the need for clinicians and patients to track medical knowledge. For example, a clinician would like to be notified with relevant search results whenever there is an important new press release and it is likely that his or her patients will make inquiries about this topic. Based on multiple studies in the UK healthcare domain, Hinze et al. (2006) identify several user requirements for alerting in health digital libraries and go on to describe system-level implementation details. Although these particular studies (and others such as Collins et al., 2005) are specific to their unique context, they raise important architectural and temporal issues for designing and evaluating awareness mechanisms for digital libraries. We contextualize and discuss our findings with respect to these issues toward the end of the paper.

Based on the literature review above, there is clearly a need to support awareness of digital library resources and explore their potential to support specific user needs (Crabtree, Twidale, O'Brien, & Nichols, 1997). In this paper, we are concerned with supporting awareness of digital library resources. Particularly for scholarly digital libraries, we are interested in providing awareness mechanisms for users to stay cognizant of new publication events, similar to how Collins et al. (2005) use information feeds as awareness mechanisms for new scientific information.

3. Citeseer: history and background

Our study context is CiteSeer (Giles, Bollacker, & Lawrence, 1998): a free public resource providing access to the full-text of nearly 700,000 academic science papers and over 10 million citations in the computer and information science domain. CiteSeer currently receives over approximately 1.5 million hits a day and is accessed by 150 countries and over a million unique machines monthly. CiteSeer was created by Kurt Bollacker, Lee Giles, and Steve Lawrence in 1997–1998 at NEC Research Institute to provide a comprehensive and accessible digital library and search engine for the computer and information science research community that would be focused on citation linking. The query "CiteSeer" returns millions of unique documents from the popular Internet search engines Google and Yahoo and is widely indexed by both. CiteSeer is frequently cited as a search service that has greatly improved communication and progress in computer science research. It is currently hosted and maintained by the College of Information Sciences and Technology at The Pennsylvania State University.

CiteSeer consists of three basic components: a focused crawler or harvester, the document archive and specialized index, and the query interface. The focused spider or harvester crawls the web for relevant documents in PDF and Postscript formats. After filtering crawled documents for academic documents and possible duplicates, these are then indexed using autonomous citation indexing, which automatically links references in research articles to facilitate navigation and evaluation. Automatic extraction of the context of citations allows researchers to determine the contributions of a given research article quickly and easily; and several advanced methods are employed to locate related research based on citations, text, and usage information. CiteSeer is a full text search engine with an interface that permits search by document, numbers of citations, or by fielded searching, not currently possible on general-purpose web search engines.

It is traditional practice in the computer and information science community to make research documents available at the time they are first written through technical reports series managed by various laboratories and academic departments. More recently, this practice has been transferred to the World Wide Web (Goodrum, McCain, Lawrence, & Giles, 2001). CiteSeer actively and automatically harvests these documents and automatically builds searchable and indexable collections, promoting creative scientific discovery and reuse within the computer and information science community. Even though search engines such as Google actively index CiteSeer, users come to the CiteSeer engine for unique information such as citation counts and domain dependent citation links not provided by Google or Google Scholar.

We now transition to the empirical part of our paper. The three studies we conducted with CiteSeer users follow in subsequent sections (Sections 4–6, respectively).

4. Requirements elicitation: survey of CiteSeer users

Our goal is to enhance CiteSeer as a *collaboratory*: a center without walls, in which 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 (Wulf, 1993). Our premise is that direct collaboration between peers in a scientific community around existing digital library resources would lead to more meaningful and long-term collaborative endeavors and scientific outcomes.

As part of our requirements gathering process for the CiteSeer collaboratory, we conducted an online survey and follow-up email interviews with CiteSeer users. The objective of the survey was to gain insight into the kinds of activities CiteSeer users would like to collaborate on and identify possible socio-technical issues during such collaboration.

4.1. Recruitment and participants

The survey was made available on CiteSeer's web site. Participants were CiteSeer users willing to take the online survey. The opportunity sample, based on self-selection, was the only realistic sampling procedure available to us. There is no feasible way to randomly identify or select a sample of CiteSeer users, because the population of users is not enumerated anywhere. One becomes a user merely by accessing CiteSeer services.

The survey link was placed in multiple locations on CiteSeer's web site, and read, "Help us improve CiteSeer. Take a survey". Clicking on the survey link would direct users to an informed consent web page and upon acceptance of the consent form, users would be redirected to the survey questions. No compensation was provided to survey responders. The results reported here are based on the administered survey for one month (mid November–December 2005).

4.2. Survey design

We designed the survey asking 29 questions (12 of which were multi-part) organized into four broad sections:

- *Professional interaction*. Seven questions (five of which were multi-part), related to how often CiteSeer users collaborated face-to-face and remotely, how they would like to collaborate with other CiteSeer users, and what issues they might face in online collaboration.
- *CiteSeer use*. Seven questions (two of which were multi-part), related to how often CiteSeer users used the search engine, the nature of CiteSeer queries, and whether or not the use of CiteSeer led to collaboration.
- *Comparison of search engines with CiteSeer*. Six questions (five of which were multi-part), related to the use of CiteSeer with other academic search engines such as ACM Digital Library, IEEE Xplore, and Google Scholar.
- *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, frequency of CiteSeer use) and ratings on 7-point Likert scales (e.g., frequency of use for a specific CiteSeer feature on a scale of "Never" to "Very often"); free-text opportunities were also provided (e.g., academic background). Based on pilot testing, the survey required approximately 10–15 min to complete.

4.3. Data collection and analysis

The number of participants who responded to the survey was 562. Some respondents skipped one or more questions. The last survey question asked participants if they were willing to be interviewed via email. We contacted 126 participants and got responses from 41 participants. The second to last question in the survey asked for any type of feedback from participants related to CiteSeer; 129 participants responded. We analyzed these responses for any feedback related to the interview questions.

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

4.4. Results

Seventy-five percent of the respondents had a background in computer science. The participants represented a relatively core group of CiteSeer users. Their mean use of CiteSeer was 3.5 years. Thrity-nine percent downloaded more than 100 papers through CiteSeer. Sixty-nine percent of the respondents indicated that they use CiteSeer at least once or twice per week. Results from this survey are summarized in Table 1. In this paper, as we are interested in awareness mechanisms for CiteSeer, we briefly expand on the last two findings.

Table 1
Summary of survey results based on requirements gathering from CiteSeer users

Requirements	Examples
Visualize social networks to identify scholarly communities of interest	Social networks based on shared queries or citations.
Provide online collaborative tool support for upstream stages of scientific collaboration	Brainstorming activities; topic-based discussions
Support awareness to stay cognizant of new publication events in the digital library	Citations to one's papers; new papers; related paper to one's papers
Use notification systems to convey awareness that is peripheral to users' primary task	RSS feeds

This survey study shows that supporting awareness of new publication events (e.g., "new articles", "what's new in my field", "what other researchers are thinking or focusing on") is desirable. Part of the challenge in supporting awareness is knowing *how* to convey it effectively. Survey results suggest that users preferred RSS feeds as notification systems. Notification systems are defined as interfaces that are typically used in a divided-attention, multi-tasking situation, attempting to deliver current, valued information through a variety of platforms and modes in an efficient and effective manner (McCrickard et al., 2003). They 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. (NB: Although other terminology in concurrent research has been used such as "alerting services" (Buchanan & Hinze, 2005), we will use "notification systems" in this paper.)

The design rationale for using notification systems is based on at least two reasons. First, awareness of new publication events is not the primary task of users, hence 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. According to McCrickard and Chewar (2003), the success of a notification system hinges on accurately supporting attention allocation between tasks, while simultaneously enabling utility through access to additional information.

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.

5. Prototype evaluation: assessing user preferences for RSS feeds

The first study shows that CiteSeer users want to use RSS feeds as notification systems to stay aware of new publication events such as citations to one's papers, new papers in one's research area, and related papers to one's papers. When designing content for notification systems, an important issue to consider is what awareness information will trigger the user's interest and encourage him/her to follow-up on the notification (see detailed discussion on *comprehension* and *reaction* in McCrickard et al., 2003).

The awareness information in an RSS feed should provide a useful and relevant summary of a new publication event through concise, but meaningful content (Hylton, Rosson, Carroll, & Ganoe, 2005). This is because there is a tension between wanting to display a lot of information and not wanting to take up precious screen real estate or the user's time (Cadiz, Venolia, Jancke, & Gupta, 2002). Our second study investigates different content designs of awareness information for various RSS feeds.

5.1. Experimental design

We conducted a formative evaluation to assess the feasibility of three content design candidates each for three different types of RSS feeds. The first type of RSS feed was "Citations", which lists papers that cite any of your papers. The second type of RSS feed was "New papers", which lists papers in your areas of research interest. Research areas are indicated by keywords. The third type of RSS feed was "Related papers", which lists papers that are related to any of your papers. Relationship is established based on co-citation: related papers are those cited by papers citing your paper. For example, paper "A" is related to your paper "B" if some collection of papers (anchor papers) cite both "A" and "B".

These three types of RSS feeds were chosen, because they most directly represented the kinds of publication events that CiteSeer users wanted to track, based on the previous study. Also, these feeds can be easily implemented in CiteSeer as the database already keeps track of citation networks, keyword occurrence in papers, and strength of relationship between papers.

The "Citations" feed can be presented as three different target items: (a) Reference (ref): full reference of the paper that cites your paper; (b) Reference and abstract (ref + abs): full reference of the paper that cites your paper and its abstract; or (c) Reference and context of citation (ref + context): full reference of the paper that cites your paper and the sentences in the paper that contain the citation to your paper.

The "New papers" feed can be presented as three different target items: (a) Reference (ref): full reference of the new paper; (b) Reference and abstract (ref + abs): full reference of the new paper and its abstract; or (c) Reference and keyword occurrence (ref + keywords): full reference of the new paper and the sentences in the paper that contain the research area keywords you specified for the feed.

The "Related papers" feed can be presented as three different target items: (a) Reference (ref): full reference of the related paper; (b) Reference and abstract (ref + abs): full reference of the related paper and its abstract; or (c) Reference and contexts of citation (ref + contexts): full reference of the related paper and the sentences in the anchor papers that contain the citation to your paper and the related paper.

For each feed, the three target items were chosen as a result of iterative refinement among the researchers. The researchers brainstormed initial best guesses about the most likely content that users would like to see in the various RSS feeds. Pilot studies suggested that these design choices were appropriate candidates for evaluation.

5.2. Task

We asked CiteSeer users to evaluate which of the three design candidates for each type of RSS feed was most useful and relevant. Interface screenshots of the feeds were presented to the users after which they would answer questions indicating their preferences. The feeds were real in the sense that they pointed to papers available in CiteSeer. Users could click on the screenshots and be taken to the paper on CiteSeer. HTML image maps were used to create these screenshot hyperlinks.

We used Survey Monkey (http://www.surveymonkey.com/) to evaluate the interface screenshots and collect participant responses. The interface screenshots of the RSS mock-up feeds were embedded in the survey itself, so that users could look at the design candidates and give feedback immediately. Counterbalancing was done to ensure order of the presentation of feeds and their design candidates did not affect user preferences.

Because we did not know the identity of the participants, we asked participants to role-play a well-known researcher while evaluating the different designs: Larry Page, co-founder of Google. This helped contextualize the task, as we were able to configure the RSS feeds to report citations and related papers to any of Larry Page's three famous 1998 publications ("Bringing order to the web", "The anatomy of a large-scale search engine, and "Efficient crawling through URL ordering"). New papers were reported based on three search keywords that Larry Page might find useful ("Web crawling", "Personalized web search", and "Google").

5.3. Recruitment and participants

The survey was made available on CiteSeer's web site. The hyperlink invitation to the survey read, "Give feedback on RSS feeds for document recommendations in CiteSeer". Clicking on this link would direct users to an informed consent web page. Upon acceptance of the consent form, users would be redirected to an instructions page. The instructions fully explained the task and experimental design (e.g., details of how related papers are being recommended). After reading the instructions, users clicked on a link to continue onwards for viewing the interface screenshots of the feeds and answering the survey questions. No compensation was provided to participants.

5.4. Measures

We wanted to understand why certain design candidates were better than others. We asked users to pick the most useful and relevant design candidate for each feed description and explain their rationale. Pilot studies indicated that users also wanted to specify which design candidate was the least useful and relevant. We included this in the questionnaire as well. Users were also given free-text opportunities to suggest any relevant changes or enhancements to the existing design candidates.

Users were asked questions about their use of CiteSeer (e.g., number of years they have been using CiteSeer) and RSS technology (e.g., number of times they check their RSS feeds). Several demographic questions were also asked (e.g., academic qualification). The survey required approximately 20 min to complete.

5.5. Overall results

The results in this paper are based on the administered survey for four weeks. A total of 57 participants took the survey. Nine participants did not indicate any preference for the feed design candidates; these responses were not considered in the analysis, so we report the results based on 48 survey respondents. Some of these respondents skipped one or more questions.

Of the responses we received, 39% were graduate students and 22% were university professors. Males (89%) outnumbered females. The sample as a whole was relatively highly educated, with 40% having a master's degree and 40% having a doctorate degree. The survey respondents represented a relatively core group of CiteSeer users. Their mean use of CiteSeer was 3.6 years. Their mean use of RSS technology was 1.9 years. 42% of the respondents said they check their RSS feeds once or twice per day. Forty-two percent said they check their feeds several times a day.

Fig. 1 shows user preferences for the different design candidates for each of the three types of RSS feeds in CiteSeer. The "Like" and "Dislike" categories indicate what users found as the most and least useful and relevant design candidates respectively. For each type of RSS feed, chi-square (X^2) tests showed a statistically significant relationship between user preferences ("Like", "Dislike") and the three design candidates.

For citations ("Citations"), most users (57%) preferred the "reference & context of citation" design candidate (X^2 (2, $N = 59$)=19.77, $p \leqslant 0.001$). For new papers ("New"), majority of users (64%) preferred the "reference and abstract" design candidate (X^2 (2, $N = 77$)=23.22, $p \leqslant 0.001$). For related papers ("Related"), most users (54%) preferred the "reference & abstract" design candidate (X^2 (2, $N = 73$)=19.03, $p \leqslant 0.001$). For all the three types of feeds, users clearly thought that just the "reference" design candidate was the least useful and relevant.

5.6. "Citations" feed

Users are interested in citations to their papers primarily because of how they are cited. One user said: "Helps to know as early as possible the interest of the citation". Users expressed the view that the context of citation may be the most useful and relevant piece of information in a citing paper: "A citation is useful due to its context of citation. It saves the step of searching for the context of citation". One user even compared citations to a blog, suggesting that discussions around a blog entry are in fact the most valuable contributions: "It's like a blog. The paper is a post and the context of citation is the comments".

Most users did not like the reference candidate, because it provides "no information about what the paper is about." The reference and abstract candidate was considered by users to be "too general and therefore less informative."

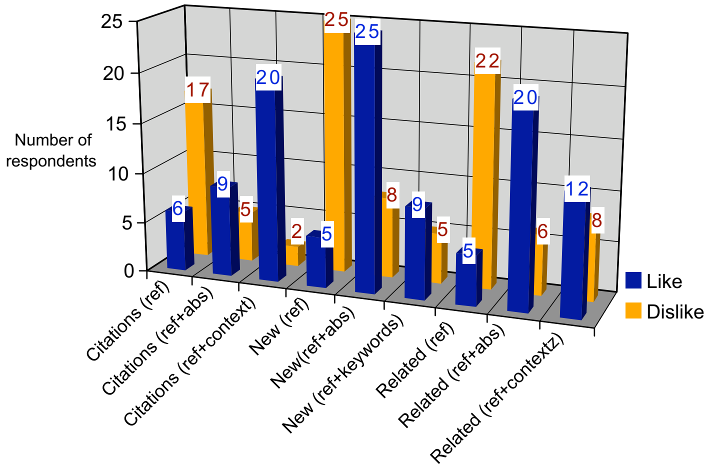

Fig. 1. User preferences for RSS content design mock-ups in CiteSeer.

5.7. “New papers” feed

Most users preferred seeing the abstract of a new paper, because it was the primary criterion to decide whether or not to read the paper itself. One user said: “It gives a good idea what the paper is about and whether it is worth reading.” Users wanted to know “what the paper is about” before investing too much time and they agreed that “a quick abstract is really useful”.

Users thought that the reference-only candidate contained “too little information”. As for the keywords, users thought they were “noise” as the “frequency of the keyword(s) makes no contextual sense”. One user explained why this was the case, suggesting that keywords can be misleading: “I believe that the keywords are in the document, I do not need proof of that. Sometimes keywords are occurring, so often that this view is worthless”.

5.8. “Related papers” feed

Users preferred to see the abstracts for related papers for similar reasons as for the new papers, because they wanted to “immediately decide whether it (the related paper) is worth reading”. As before, users felt that just the reference did not provide enough information and that the abstract saves the most time in deciding about further investing time. One user said: “The title says nothing. I’m not going to click on every link to see if it is worth reading, I want to know immediately”. As for the contexts of citation for related papers, one user said that even though similarity might be reflected through sentences in the context, the related paper is only of interest based on its content: “I expect ‘to be related’ being based on paper-level (rather) than on sentence-level”.

5.9. Discussion

This study shows that users prefer feeds that place target items in query-relevant contexts, and that preferred context varies with type of publication event. For the citations publication event, the query-relevant context for users seems to be the context of citation. For new and related papers, the abstract seems to be the preferred target item that provides the most query-relevant context to users.

The fact that what participants identified as query-relevant context varies with the type of publication event makes sense when the use of feeds is considered. For different types of publication events, users have different goals. Their preference is dependent on the contextual information that provides the most useful and relevant cues for their goals.

For the citations publication event, the primary goal of users was generally to know how their paper was cited and not what the citing paper was about (although this may be a secondary goal). Hence, the context of citation provided the most query-relevant context to users. This suggests that people are very interested in what others think of their work, which again makes sense in the highly competitive and evaluative world of academic publications.

For the publication events of new and related papers, the goal of users was generally to decide whether or not they should read the paper. Reading the abstract was the most efficient and effective way to achieve this goal. This shows that CiteSeer users are also interested, though perhaps to a lesser extent, in making sense of other people’s work, and finding interesting work relevant to their own.

Our results indicate that reference-only feeds are always least preferred, regardless of the type of publication event. This implies that a reference alone does not provide enough query-relevant context to the users. When designing content for RSS feeds, it is important to know how much awareness information can provide the query-relevant context to users. Providing too little awareness information (e.g., only the reference) is not sufficient to trigger a user’s interest. In this sense, including the reference in the content design seems to be the minimum awareness information required to provide this query-relevant context. Although we did not ask users to rate design candidates without the reference, no one indicated their dislike for the reference when other contextual information was also provided. We suspect that the reference is necessary, because it provides useful and relevant information about a publication event, such as the authors, publication venue, and year of publication.

Users indicated that they would like to customize the target items in RSS feeds. For example, one user said the following for the citations feed: "I would prefer both the abstract and context (of citation) in an RSS feed". Users also suggested feed ideas for other publication events, such as tracking papers written by specific authors, institutions, or research groups. We are considering these enhancements as immediate future work.

6. Naturalistic study: users collaborating with rss feeds

Based on prototype evaluation, the second study determined preferences for various feed content designs. We now wanted to understand how users would integrate the feeds in the context of their broader, everyday activities over several days. The third study evaluates CiteSeer RSS feeds in a naturalistic collaborative setting based on the feed preferences determined from the second study.

We conducted two evaluations as part of the third study. For the first, *initial* evaluation, three teams comprising two users each were asked to collaborate on a shared task of writing a technical opinion piece remotely over the duration of four days. Although we found interesting and useful results, we learned three lessons from this initial evaluation: (1) The duration of four days was too brief for participants to adequately and authentically complete the task; (2) Two team members were not enough to complete the task as a group (e.g., one member drops out); and (3) Fully distributed collaboration (no face-to-face interaction) led to lack of initial common ground, which subsequently caused failure in group dynamics.

Learning from the above lessons, we ran a second, follow-up evaluation as part of the third study, which was conducted over one week in which each of the three teams comprised three group members. We expected that the follow-up evaluation would validate and build on the results of the initial evaluation – this was indeed the case. Following are experimental details of both the initial and follow-up evaluations; differences between the two evaluations are explicitly mentioned throughout the description.

6.1. Experimental setup

Users worked on the shared task in a collaborative environment called BRIDGE (Basic Resources for Integrated Distributed Group Environments; http://bridgetools.sourceforge.net; Ganoe et al., 2003). The BRIDGE Java-based client supports shared editing of documents through replicated objects. Replicated objects are objects that are retrieved by multiple collaborating sessions and whose state is kept synchronized on all clients and a server when any replica is changed. The following collaborative tools were provided in BRIDGE: persistent chat tool (for ad hoc communication), brainstorming space (to develop ideas), concept map (to visualize ideas), and opinion piece (for the final product). The brainstorming space and opinion piece were wiki-based. All these tools supported synchronous and asynchronous collaboration.

6.2. Task

Each team was asked to work in a distributed fashion using BRIDGE. Their task was to write an opinion piece related to computer science education, specifically on introducing and teaching software programming to new computer science students. This task was chosen, because recently, in *Communications of the ACM*, opinion pieces have been published on this topic (Hu, 2005; Martin, 2006). The teams were told that they would be assessed on the novelty of their opinion piece.

As guidance to the teams, they were given a foundation paper by Westfall (2001), and were asked to expand, critique, or base their opinion piece on this paper. Users were told they would get other RSS feeds for papers while they were working on their opinion piece. Three RSS feeds were sent to the team. The first one was a "new paper" feed, pointing to Hu's (2005) paper on CiteSeer. The second one was a "citations" feed, pointing to Martin's (2006) paper that cited Hu's (2005) paper. The third one was a "related paper" feed, pointing to Martin and Kuhn's (2006) paper as it was related to Martin's (2006) previous paper. The content design for all these feeds were based on the preferences elicited by users in the second study. These feeds were sent in the order mentioned above. Each feed was sent on a different day by the experimenter.

6.3. Recruitment and participants

Team members were recruited from the graduate student population of the College of Information Sciences and Technology at The Pennsylvania State University. Graduate students were chosen, because they represented the majority of users who use CiteSeer (based on our results from the first CiteSeer survey). Users willing to take part in the study were recruited. The teams were formed opportunistically based on the availability of participants to volunteer for the study.

For the initial evaluation, we refer to the three teams (and their members) as Team A (members A1 and A2), Team B (members B1 and B2), and Team C (members C1 and C2). For the follow-up evaluation, we refer to the three teams as Team X (members X1, X2, and X3), Team Y (members Y1, Y2, and Y3), and Team Z (members Z1, Z2, and Z3).

Teams A, B, and C (initial evaluation) were given four days to complete the task whereas Teams X, Y, and Z (follow-up evaluation) were given one week. Members of Teams A, B, and C collaborated in a fully distributed manner. For this initial evaluation, the experimenter emailed instructions to each team member, met with them to make sure they understand the instructions, and demonstrated how BRIDGE works.

For Teams X, Y, and Z, members of each team were invited to an initial kickoff face-to-face meeting (approximately 25 min). Each member was introduced to the others. The experimenter explained the task, demonstrated the use of BRIDGE, and clarified any issues related to the task or technology. The experimenter emphasized that the collaboration should be carried out only through BRIDGE. Approximately 5 min toward the end of this meeting were allotted for social grounding. Based on the initial evaluation, we considered it essential to provide group members in the follow-up evaluation with a face-to-face opportunity to establish initial common ground and strategize about their subsequent collaboration.

6.4. Measures

After completion of the task, we conducted semi-structured interviews with the team members that lasted approximately 20 min each. We asked questions related to how they reacted to the RSS feeds when they first saw them, what actions they took in terms of collaboration after reading the feeds, what they liked and disliked about the feeds, and how the feeds affected their collaborative task in BRIDGE.

The analysis of the interview data for each team was done using the general analytic strategy of developing a case description (Yin, 2003). We summarized and compared findings from each team with the others with respect how the feeds were integrated in the context of the underlying task, thus providing a rich cross-case interpretation. We also culled and collated user quotes from the interviews by identifying appropriate instances that helped to triangulate our collaborative interpretation of the data.

We analyzed each team's collaborative behavior in BRIDGE. Team members communicated with each other using a text chat tool. In our results, we present instances from these chat logs related to the CiteSeer RSS feeds.

6.5. Results and discussion

All teams completed their task with the exception of Team C. After reading the instructions for the task, user C1 declined to do the task, because he felt that he did not have enough academic background to write an opinion piece on teaching programming. User A2 did not contribute to the task, as he was unable to set up the RSS feeds on his computer.

All users who actively used the RSS feeds indicated that the feeds were related to the task and helped them articulate their opinion piece. User Z1 commented: "All three (feeds) were useful for the discussion...It gave us enough feedback to start the discussion." Although users said that the feeds were relevant to the task, their main concern was whether the feeds would be relevant beyond the context of the task. For example, user A1 said: "It was obvious they (the task and the feeds) were related. That was a good way of sharing papers and information. It'd be awesome, like, if I got fed, like, open-source stuff (this user's research area)." After the experimenter explained the ultimate goal of the RSS feeds is to personalize them for individual users, participants could relate to how these feeds would be useful to them in the context of their everyday activities.

Most users indicated that they did not immediately react to the RSS feeds when they received them. When users had the opportunity to work on the task, they followed the feed links and read the papers. User B1 said: "When I needed to write the paper, I downloaded it and read it". This was understandable, as the task itself did not demand critical attention. Users did however take notice of the feeds once they received them, even if they did not click on the feed link. For example, user X1 said: "I did not necessarily read them as I got them, but I did recognize they were there ... If it had anything new to add to our conversation, we would discuss it". In another instance, user A1 said that she read the feed content, which was sufficient to inform her that the information was useful for later: "I read it (content) from the feed . . . that was enough to make me go . . . oh okay . . . enough information (for later)".

When users set aside time to work on the shared task, they did look at the RSS feeds and also logged into BRIDGE to communicate, coordinate, and share ideas with their team members. User B2 said: "I think its kind of iterative, because I first read one paper and then I got to BRIDGE. I talked with [user B1], we exchanged some ideas." User Y3 indicated that she actually uploaded a link on BRIDGE to one of the papers from the feeds so that other team members could have a follow-on discussion. Such episodes illustrate that the RSS feeds were an integral part of the shared task.

All users indicated that they assumed their team members received the RSS feeds just as they did. In some cases, team members used the feeds as planning tools to decompose the shared task. For example, in Team B, user B1 left the following chat message for his team member: "I see there are two (related) papers from the CiteSeer feeds, but I have not read (them) yet. Let's write whatever we think about the three papers (including the foundation paper) in the 'final opinion piece' object, and we merge our ideas on Saturday and Sunday". User B2 later said in the interview that he used the feeds not just for planning, but also as milestones to monitor team progress: "I think its something like a checkpoint where you communicate with your partner to see what's going on". User Z1 from Team Z also expressed similar thoughts, saying that the feeds helped him to "organize (his) ideas". Using feeds as planning tools can be useful during scientific collaboration to coordinate shared tasks.

We also asked users about what they disliked about the RSS feeds and other improvements. Two users reiterated the problem we identified from our second study; users were not sure of how the papers were selected in the RSS feeds. User C2 said: "I did not know on what basis subset was made. It was not clear why those two were (related)". This emphasizes that it is not only important to select and present useful and relevant papers, but also convey to the users how the papers were chosen. Conveying such selection criteria seems to be an important requirement while designing such RSS feeds.

Users, in general, acknowledged the usefulness of the feeds. For example, user X2's comment echoed similar reactions of others toward the personalization of feeds: "The idea that I get the papers that are very relevant to what I am doing, that inform me what I'm doing now – I think it's very beneficial". One user also said that beyond the context of the shared task, one may need more feed content than what is provided now in order to better decide which papers to read. User B2 said: "I just went directly (to CiteSeer for this task) . . . If there are five or more feeds, so I might expect more information (to help me decide which paper to read)". This suggests that more meta-information in the feed content may be required as the number of feeds for an individual increases. This is not surprising; as the number of papers scale, one wants to know more information to assess the usefulness and relevance of those papers.

Although we ran a limited number of participants in our third study – due to the complex logistics of carrying out naturalistic studies of distributed collaboration over fairly extended periods of time – we observed explicit patterns of user behavior toward the feeds and how users integrated them in the context of an authentic task. Overall, users found the RSS feeds to be useful for the underlying task that was assigned to them. Many alluded to the potential benefits of the feeds during realistic settings when users would be engaged in their broader, everyday activities. This is encouraging as we plan to implement and deploy actual feeds for CiteSeer based on feedback from our user studies.

7. General issues and future work

The three cumulative studies represent an iterative design-based research cycle to study the use of digital libraries in the context of users' broader activities. We were able to collect data from multiple sources of research investigation (requirements elicitation, prototype evaluation, and naturalistic study) and triangulate our findings.

In addition to immediate design improvements to CiteSeer's awareness mechanisms, the three user studies point to several higher-level issues related to awareness in scholarly communities. Specifically, the user studies raise four practical considerations related to implementing awareness mechanisms: sensemaking in communities, liveness of publication events, personalization of feeds, and architectural issues.

7.1. Awareness as sensemaking in scientific communities

Scientific communities have traditionally formed around key intellectual resources such as collections of books, or special equipment such as cyclotrons (Wellman, 1999). 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 colocated with their shared resources and with one another. Since its inception, the World Wide Web has changed the ways scientists search for and access intellectual resources. The Web has reduced the effort of finding previous work, contacting colleagues, consulting with peers and mentors, and disseminating scientific contributions (Kiesler, 1997). To this end, digital libraries have provided dramatic improvements over traditional libraries by enhancing searching, browsing, and filtering capabilities.

Although digital libraries are indeed repositories for information search and retrieval, they are also services that attract people and facilitate the formation of scholarly communities around their collective resources. Users are doing more than visiting a web site; they are making sense of their search activities, building and sharing knowledge, and collaborating with other community members. They are in fact participating in online scientific *communities of practice*. A community of practice involves a set of socially defined ways of doing things in a specific domain (e.g., computer and information science): a set of common approaches and shared standards that create a basis for action, communication, problem solving, performance, and accountability (Wenger, 1998). A community of practice caters for professional practice and development. It provides access to a network of community members in various types of social and professional relationships with each other.

The CiteSeer population can be seen as an untapped community of practice. CiteSeer users have the basic characteristics of a community of practice – domain of knowledge, community of people, and shared practice – but they do not have any online mechanism that allows them to see, stay aware of, or interact with one another. The most significant result from our CiteSeer findings is indeed that users want to and can collaborate around the intellectual resources of a digital library in ways similar to that in an online community of practice. A corollary to this result is that scientific communities around digital libraries can be better supported by tools that reinforce their identity as a community and provide an incubating environment to collaborate with others in the community.

Membership in a community of practice can be utilized to enhance the sensemaking process at the individual level. To achieve this, members need to be aware of each other's activities, who is doing what, when a particular piece of digital information has been updated, and so on. Changing search within a community from a static action (e.g., user searching on a portal) into something that occurs live over time provides opportunities to make sense of the activity and enhance collaboration within that community. Indeed, feed-based notification systems facilitate such sensemaking through their aggregated search results over time.

By broadening search from a short-term ephemeral task into an embedded long-term activity, we can better understand what is going on in that community. Activity awareness considers how one stays aware of long-term efforts directed at specific goals and objectives to promote informed action and reaction (Carroll, Rosson, Convertino, & Ganoe, 2006). Within a scholarly community, this activity information includes who is currently active in the field, what they are working on, how it relates to other work in the field and who else is interested in this work.

Sensemaking of the activity information can be aggregated from the temporal information in the feeds. This is consistent with Reddy, Dourish, and Pratt's (2006) research that highlights the opportunity of leveraging temporal information to understand patterns of former actions and expectations about future activities. While a manual CiteSeer search on "tactile interfaces" might give you all the instances of that phrase in that site's database at that point in time and who wrote them, aggregating what occurred over time with who created and/or is seeking those resources can answer questions like: Is this topic active now or when was it active in the past? Who else is publishing or even searching in this area? Are they active now and/or when were they active in the past? Answering such questions can also provide reliable indices into how users have integrated feeds in the context of their broader, everyday activities.

7.2. "Liveness" of publication events

Providing access to information is of little use unless the information is timely (Collins et al., 2005). We refer to this issue as "liveness" of the publication events. Notification of new publication events, such as the acquisition of new papers that cite one's prior papers, should occur as close as possible to the time of the event and should be distributed as evenly as possible over time. This is because of two reasons. First, receiving timely notifications may be critical to users' current activity in hand, such as writing and submitting a conference paper by a specific deadline. Such a scenario would be similar to the shared task in our third user study.

Second, at the level of design and implementation, it is important to reduce the occurrence of information spikes. Data acquisition processes must be run frequently enough to ensure that content is up-to-date and that the rate of data acquisition is smooth. For instance, document and citation data is currently updated in CiteSeer through expensive batch processes that are run at most every two weeks. This means that notifications of content acquisition events cannot occur more frequently than every two weeks without artificially delaying notifications, and that event notifications either come in batches or fall out of date. A better infrastructure for designing awareness mechanisms will either update continuously or if batch processes are required, very frequently, thus providing a stream of notifications rather than bulk distributions.

The issue of liveness for publication events only takes into account the temporal nature of the feeds (i.e., are they current and up-to-date?). But another related issue is the effect of temporal context on users' routine endeavors. For instance, considering an example from medical work (Reddy et al., 2006), does the temporal context of information in the feeds allow users to better coordinate and accomplish their everyday activities? This question can be extended, in other words, to scientists in general for investigating when users would like to receive notifications in a way that is consistent with their everyday activities. Intelligent interruption management (Adamczyk, Iqbal, & Bailey, 2005) can address such questions; we leave this to future work.

7.3. Personalization of feed-based notification systems

RSS feeds are one way users can stay aware of new publication events and integrate such notifications as part of their everyday collaborative endeavors. Based on results from our third user study, we speculate that users would not only find RSS feeds useful for staying aware of new publication events on an individual level, but could also use the feeds as a point of common ground in communication and as an anchor for collaborative interaction. These shared feeds could foster collective action, allow for the sharing of resources in their respective community of practice, and aid in the planning of more focused tasks such as jointly writing papers.

In our third study, users were presented with generic feeds that were common to all members of each team. Realistically, these feeds would be personalized for each user (e.g., citations based on papers that user has written). It is important to note that such personalization entails a design tradeoff such that it can possibly diminish common ground, as all users are not subscribing to the exact same feeds. A compromising solution, in some cases, might be to consider personalization for a group rather than for individuals (e.g., new papers based on interests of a research lab).

Personalization requires generic feeds to be specialized. At a design level, awareness features are attached to specific user preferences in order to reduce the size of data channels while providing content that is appropriate and relevant to users. A typical RSS approach that is common for blogs and news sites, for instance, is to provide a global channel for all entries such that subscribers may filter the results according to preferences supplied within a client-side application; however, this approach is not applicable for large-scale data providers for various reasons (Damianos, Wohlever, Kozierok, & Ponte, 2003). For example, the CiteSeer archive currently holds over 750,000 documents and 10 million citations and the corpus at times grows by as many as 20,000 documents and 50,000 citations in a week. Global RSS channels would require very large downloads by clients, causing unacceptably slow performance and a great waste of bandwidth. Moving user-specified content filters to the server solves both of these problems with a reasonably acceptable cost of increased server complexity.

7.4. Architectural considerations

The current CiteSeer architecture does not support the implementation of RSS feeds. Work is underway to develop an enhanced version of CiteSeer, named CiteSeer^X (Next Generation CiteSeer), which will support awareness mechanisms architecturally based on insights from the present user studies. One of the major enhancements of the new architecture is a clean separation of data ingestion processes from the core service environment that handles user queries (see Fig. 2). CiteSeer acquires data from external resources such as the Internet and other digital libraries. New documents are found during routine web crawling or as the result of user submissions. Upon arrival, these documents are fed into a configurable information extraction workflow which consists of an array of web services that encapsulate various extraction algorithms and a business process execution language (BPEL) engine that routes documents to the appropriate services. Information to be extracted from research papers includes titles, authors, author affiliations and contact information, abstracts, citations, and acknowledgements, among other data items. The result is an XML record that encapsulates all metadata for each paper. These XML metadata records are then passed to the core service applications via a Java Messaging Service (JMS) implementation for integration into the public CiteSeer^X service. The ingestion system is designed for continuous, on-line operation.

At this point, particular metadata items are stored in a common database and various content is indexed as required. When records are fully ingested, the original XML data records are then passed over a separate JMS channel to a web application that handles user account information. There, the XML is matched against user-specified filters in order to generate notifications. When matches are found, the appropriate user accounts are updated with event data, generating RSS-style notifications. This system is analogous to digital library alerting systems proposed by Hinze et al. (2006) who also make use of automatically derived data to facilitate the creation of notification filters, and to the local alerting architecture proposed by Buchanan and Hinze (2005) and implemented in the Greenstone digital library (Witten, Boddie, Bainbridge, & McNab, 2000). Unlike the work

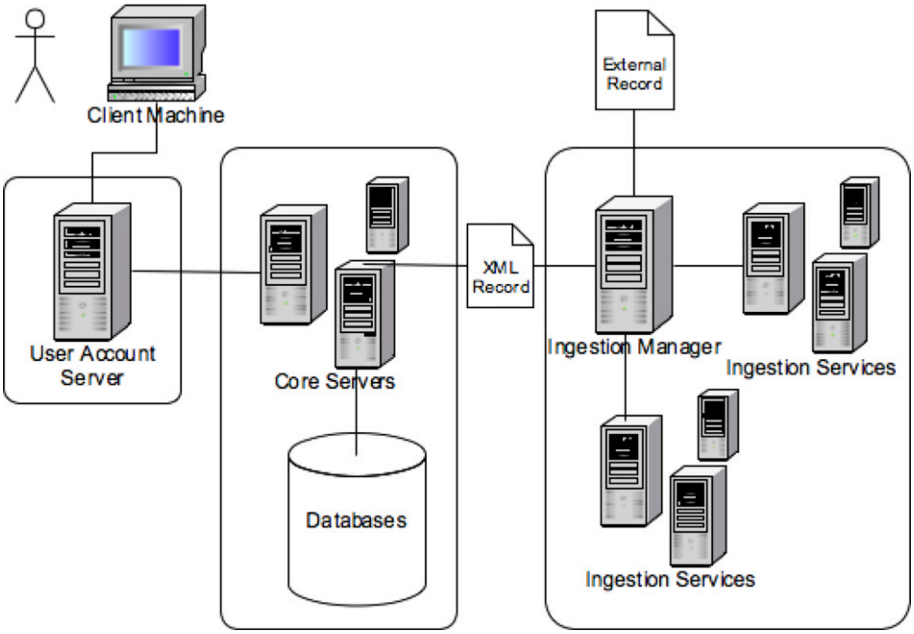

Fig. 2. The CiteSeer^X architecture maintains clear separation between ingestion services and the core service environment. Metadata regarding new content objects are passed to the core servers as XML records and is passed on to user account-based application servers for event notification. The core servers may also generate events that will be passed on to application servers.

resulting in the Greenstone alerting service, our work does not propose a generic protocol for digital library alerts, but the architecture is flexible enough to allow arbitrary interfaces into the system in order to facilitate the adoption of future alerting standards.

The XML structure of content records allows the creation of a powerful and extensible filtering language to enable a rich set of options to users. Content may be tracked in several automatically parsed document text fields such as document bodies (the standard search), titles, abstracts, publication venues, document headers, author affiliations, acknowledgement text, and other fields. This enables users to track not only topics that are of interest, but also allows institutional tracking such as documents that are published by authors from a specific institution or papers that acknowledge a certain funding source (e.g., Giles & Councill, 2004). In addition, the core servers will propagate any data changes to the account server such as when new citation matches are found, a document's metadata is changed, or new documents are found to be similar to other documents. Together, these notification sources provide a powerful set of hooks into the dynamic system data, which users can employ to craft personalized awareness strategies. Upon deployment, these facilities will be refined through further user studies.

8. Conclusion

In this paper, we presented three user studies to investigate the design feasibility and effectiveness of awareness mechanisms for scholarly digital libraries. From an initial requirements elicitation survey, users indicated that they would like to stay aware of CiteSeer's intellectual resources. Specifically, CiteSeer users wanted to use RSS feeds as notification systems to stay aware of new publication events. We presented CiteSeer users with RSS mock-up prototypes to assess their preferences for various types of publication event feeds. This second study showed that users prefer feeds that place target items in query-relevant contexts, and that preferred context varies with type of publication event. Evaluating these feeds in a naturalistic setting, we found that users integrated the feeds as part of their collaborative task, used the feeds as planning tools, and suggested future design enhancements such as having more meta-information in the feed content. Toward the end of the paper, we reflected on general issues and future work related to sensemaking in communities, temporality and liveness of information, personalization of notification systems, and architectural design. We believe our paper will be of interest to scholars in the domain of information search and retrieval who wish to enhance digital libraries as community-based and collaborative services.

Acknowledgements

This research is supported by the US National Science Foundation (NSF) grant CRI-0454052. We would like to acknowledge Levent Bolelli's assistance in setting up the studies on CiteSeer's web site. We have incorporated some feedback from an informal discussion of our user studies at the Human Computer Interaction Consortium (2007)–many thanks to Dan Russell and John Daughtry for their insights. Much to their credit, the reviewers and guest editors were helpful in refining the general discussion.

References

Adamczyk, P. D., Iqbal, S. T., & Bailey, B. P. (2005). A method, system, and tools for intelligent iterruption management. In: *Proceedings of the 4th international workshop on task models and diagrams* (pp. 123–126). Gdansk, Poland.

Adams, A., Blandford, A., Budd, D., & Bailey, N. (2005). Organisational communication and awareness: A novel solution. *Health Informatics Journal, 11*, 163–178.

Borgman, C. L. (1999). What are digital libraries? Competing visions. *Information Processing and Management, 35*, 227–243.

Borgman, C. L. (2006). What can studies of e-learning teach us about collaboration in e-research? Some findings from digital library studies. *Computer Supported Cooperative Work, 15*(4), 359–383.

Bishop, A. P., Van House, N., & Buttonfield, B. P. (Eds.). (2003). *Digital library use: Social practice in design and evaluation*. Cambridge, MA: MIT Press.

Blandford, A., & Gow, J. (2006). Digital libraries in the context of users' broader activities. *D-Lib Magazine, 12*(7/8), 2006.

Buchanan, G., & Hinze, A. (2005). A generic alerting service for digital libraries. In: *Proceedings of the joint conference on digital libraries* (pp. 131–140). Denver, Colorado, June 7–11.

Cadiz, J. J., Venolia, G. D., Jancke, G., & Gupta, A. (2002). Designing and deploying an information awareness interface. In *Proceedings of the conference on computer supported cooperative work* (pp. 314–323). New York, NY: ACM Press.

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

Collins, L. M., Mane, K. K., Martinez, M. L. B., Hussell, J. A. T., & Luce, R. E. (2005). ScienceSifter: Facilitating activity awareness in collaborative research groups through focused information feeds. In *First international conference on e-Science and grid computing* (pp. 40–47). Melbourne, Australia: IEEE Computer Society, December 5–8.

Crabtree, A., Twidale, M., O'Brien, J., & Nichols, M. (1997). Talking in the library: Implications for the design of digital libraries. In: *Proceedings of DL* (pp. 221–228).

Damianos, L., Wohlever, S., Kozierok, R., & Ponte, J. (2003). MiTAP: A case study of integrated knowledge discovery tools. In *Proceedings of the 36th international conference on system sciences* (pp. 693). Washington DC: IEEE Computer Society.

Dillon, A. (2002). Technologies of information: HCI and the digital library. In J. M. Carroll (Ed.), *Human-computer interaction in the new millennium* (pp. 457–474). New York, NY: ACM Press.

Dourish, P., & Bellotti, V. (1992). Awareness and coordination in shared workspaces. In *Proceedings of the conference on computer supported cooperative work (Toronto, Canada, October 31–November 4, 1992)* (pp. 107–113). New York, NY: ACM Press.

Farooq, U., Ganoe, C. H., Carroll, J. M., & Giles, C. L. (2007). Supporting distributed scientific collaboration: Implications for designing the CiteSeer collaboratory. In *Proceedings of the Hawaii international conference on system sciences (Waikoloa, Hawaii, January 3–6, 2007)*. IEEE Computer Society, Camera-ready version available athttp://umerfarooq.net/Publications/HICSS07-CiteseerSurvey.pdf.

Ganoe, C. H., Somervell, J. P., Neale, D. C., Isenhour, PL., Carroll, J. M., & Rosson, M. B. (2003). Classroom BRIDGE: Using collaborative public and desktop timelines to support activity awareness. In *Proceedings of the conference on user interface software technology* (pp. 21–30). New York, NY: ACM Press.

Giles, C. L., Bollacker, K., & Lawrence, S. (1998). CiteSeer: An automatic citation indexing system. In *Proceedings of the conference on digital libraries (Pittsburg, PA, June 23–26)* (pp. 89–98). New York, NY: ACM Press.

Giles, C. L., & Councill, I. G. (2004). Who gets acknowledged: Measuring scientific contributions through automatic acknowledgement indexing. *Proceedings of the National Academy of Sciences, 101*(51), 17599–17604.

Goodrum, A. A., McCain, K. W., Lawrence, S., & Giles, C. L. (2001). Scholarly publishing in the Internet age: A citation analysis of computer science literature. *Information Processing and Management, 37*, 661–675.

Gutwin, C., & Greenberg, S. (1996). Workspace awareness for groupware. In *Proceedings of the conference on human factors in computing systems (Vancouver, Canada, April 13–18, 1996)* (pp. 208–209). New York, NY: ACM Press.

Hansen, P., & Järvelin, K. (2005). Collaborative information retrieval in an information-sensitive domain. *Information Processing and Management, 41*, 1101–1119.

Hinze, A., Buchanan, G., Jung, D., & Adam, Anne. (2006). HDLalert – A healthcare DL alerting system: From user needs to implementation. *Health Informatics Journal, 12*(2), 121–135.

Hu, C. (2005). Dataless Objects Considered Harmful. *Communications of the ACM, 48*(2), 99–101.

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

Kiesler, S. (Ed.). (1997). *Culture of the internet*. Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Lagoze, C., Krafft, D. B., Payette, S., & Jesuroga, S. (2005). What is a digital library anymore, anyway? *D-Lib Magazine, 11*, 11.

Martin, F. (2006). Toy Projects Considered Harmful. *Communications of the ACM, 49*(7), 113–116.

Martin, F., & Kuhn, S. (2006). Computing in context: Integrating an embedded computing project into a course on ethical and societal issues. In *Technical symposium on computer science education (Houston, Texas, March 1–5, 2006)* (pp. 525–529). New York, NY: ACM Press.

McCrickard, D. S., & Chewar, C. M. (2003). Attuning notification design to user goals and attention costs. *Communications of the ACM, 46*(3), 67–72.

McCrickard, D. S., Chewar, C. M., Somervell, J. P., & Ndiwalana, A. (2003). A model for notifications systems evaluation – assessing user goals for multitasking activity. *ACM Transactions on Computer Human Interaction, 10*(4), 312–338.

Reddy, M. C., Dourish, P., & Pratt, W. (2006). Temporality in medical work: Time also matters. *Computer Supported Cooperative Work, 15*, 29–53.

Renda, M. E., & Straccia, U. (2005). A personalized collaborative digital library environment: A model and an application. *Information Processing and Management, 41*, 5–21.

Schmidt, K. (2002). The problem with 'awareness': Introductory remarks on 'awareness in CSCW'. *Computer Supported Cooperative Work, 11*, 285–298.

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

Wenger, E. (1998). *Communities of practice: Learning, meaning, and identity*. New York: Cambridge University Press.

Westfall, R. (2001). Hello, world considered harmful. *Communications of the ACM, 44*(10), 129–130.

Witten, I. H., Boddie, S. J., Bainbridge, D., & McNab, R. J. (2000). Greenstone: A comprehensive open-source digital library software system. In: *Proceedings of the 5th ACM Conference on Digital Libraries* (pp. 113–121).

Wulf, W. A. (1993). The collaboratory opportunity. *Science, 261*, 854–855.

Yin, R. K. (2003). *Case study research: Design and methods*. Thousand Oaks: Sage Publications.