Social Bookmarking for Scholarly Digital Libraries

Social bookmarking services have recently gained popularity among Web users. Whereas numerous studies provide a historical account of tagging systems, the authors use their analysis of a domain-specific social bookmarking service called CiteULike to reflect on two metrics for evaluating tagging behavior: tag growth and tag reuse. They examine the relationship between these two metrics and articulate design implications for enhancing social bookmarking services. The authors briefly reflect on their own work developing a social bookmarking service for CiteSeer.

Umer Farooq, Yang Song, John M. Carroll, and C. Lee Giles
Pennsylvania State University

The contemporary Web has popularized *social bookmarking services*, which let users specify keywords or tags for Web resources that they're interested in. Well-known examples include del.icio.us (http://del.icio.us) and Flickr (http://flickr.com), which let users tag Web sites and pictures, respectively.

One way to measure how effective social bookmarking services are is to analyze their tag vocabularies. Two commonly used metrics are *tag growth*, which assesses the addition of new tags to the overall tag vocabulary, and *tag reuse*, which looks at the recycling of existing tags. We examine the relationship between these two metrics and how a social bookmarking service can encourage different levels of tag growth and reuse through design.

Here, we focus on social bookmarking services for scholarly communities in which users collectively organize and tag intellectual resources. Using a case study of CiteULike (http://citeulike.org), a social bookmarking service for tagging scholarly papers, we analyzed tag growth and tag reuse over time. Our results provide design implications for developing and enhancing scholarly social bookmarking services. We also briefly reflect on our own work to develop a social bookmarking service for CiteSeer,[1] an online scholarly digital library for computer science.

1089-7801/07/$25.00 © 2007 IEEE

Table 1. Data from four social bookmarking services, compared with our CiteULike data.

Name	Purpose	Data collection
del.icio.us[4]	Collaborative tagging system for Web bookmarks	Four days (212 URLs; 19,422 bookmarks)
Flickr[5]	Photo-sharing system for users to store and tag their and others' personal photos	No time data available (25,000 users)
Dogear[6]	Social bookmarking service for a large enterprise (IBM's intranet)	Eight weeks (13,174 bookmarks; 686 users)
MovieLens[3]	Movie recommender system that also lets users tag their favorite movies	Approximately one month (3,263 tags; 635 users)
CiteULike	Social bookmarking service for sharing, storing, and organizing scholarly papers	More than two years (2,011 users; 9,623 papers; 6,527 tags)

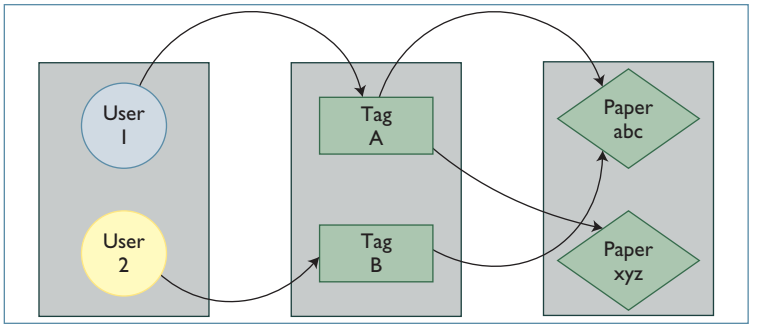

Figure 1. Anatomy of a tag application in CiteULike. User 1 (blue) has two applications and user 2 (yellow) has one tag application.

Social Bookmarking Anatomy

The basic unit of information in a social bookmarking service comprises three elements in a triple, represented as `(user, resource, tag)`.[2] Adapting terminology from previous work,[3] this triple is called a *tag application* (in which a user applies a tag to a resource; in some cases, it's also called a *tag post*). The combination of elements in a tag application is unique — that is, if a user tags a paper twice with the same tag, it counts as only one tag application.

Resources can mean different things for different social bookmarking services. With del.icio.us, for example, the resource is a Web site; with CiteULike, it's a scholarly paper.

Adapting social bookmarking's schematic depiction from Ciro Cattuto's work,[2] Figure 1 illustrates the schema for tag applications in CiteULike. This example has three tag applications: `(user 1, paper "abc," tag "A")`, `(user 1, paper "xyz," tag "A")`, and `(user 2, paper "abc," tag "B")`.

We analyze CiteULike's data around the `(user, resource, tag)` elements in the tag application. We compare and contrast our results, in general, with four other social bookmarking analyses.[3–6] Table 1 briefly lists each one's purpose and how much data the researchers collected while analyzing that service; we also include the CiteULike data set we analyze in this article.

CiteULike Overview

CiteULike is a free online social bookmarking service that lets researchers share, store, and organize information about scholarly papers. Users can add links to papers on CiteULike to their own online collections and import references from other scholarly digital libraries (Figure 2a). For example, users can link to an IEEE or CiteSeer paper in their personal CiteULike collection. The service also provides additional information about the paper, such as all users' tags for that paper and the BibTeX entry.

Adding papers to a personal collection and tagging them is a two-stem process. When users first view the link to a favorite paper, they see everyone's tags for that paper (Figure 2a). However, to add this paper as a favorite, users click on a link ("post a copy to your library") that takes them to a different tagging page (Figure 2b). On this page, users can optionally tag the paper to add it to their personal collection. Users can create new tags (by typing them in a textbox), which might overlap with existing tags others have used before, or they can select existing tags (clicking on a tag automatically adds it to the textbox), but only ones from their personal collections. Note that users don't have the option to select a tag from everyone's tag collection; if they want to do this, they have to remember the tag that others used (from when they first viewed the paper's link) and manually type it in, which we'll discuss in more depth later.

General User Activity

The analysis we describe here is based on data collected between 15 November 2004 and 13 February 2007. Although it would be interesting and useful to run our analysis on the whole CiteULike data set, because we're part of the CiteSeer research group, the underlying data set we had access to comprised only tag applications for papers in CiteSeer that CiteULike indexes.

Our data set contained a total of 32,242 tag applications, 2,011 distinct users, 9,623 distinct papers, and 6,527 distinct tags. The two most pro-

lific users had 3,883 and 634 tag applications, while 42 users had 100 or more tag applications. The two most tagged papers were both coauthored by Larry Page,[7,8] and were tagged 135 and 94 times, respectively. The five most frequently used tags were `clustering` (245), `p2p` (220), `logic` (185), `learning` (175), and `network` (175).

The average number of tag applications per paper was 3.35 (the total tag applications divided by the total number of papers). The median and modal number of tag applications per paper were 2 and 1, respectively.

The average number of tag applications per user was 16.03 (the total tag applications divided by the total users). However, the median and modal number of tag applications per user was 4 and 1, respectively. These figures are close to the ones for the MovieLens[3] analysis, which reported an average of 18 tag applications per user with a median of 3.

In MovieLens, relatively few users generated most of the tag applications, approximating a power-law distribution. CiteULike's data set is similar, with $y = 790.02x^{-1.3484}$, $R^2 = 0.9225$ (the data set included 1,921 users for a range of 1 to 55 tag applications). Figure 3 shows the relationship between the number of users and the number of tag applications.

We also computed the correlation between the number of papers each user tagged and the number of distinct tags each user generated. The correlation is high (0.944), and is thus starkly different from those of other social bookmarking services. For example, in Dogear,[6] the correlation between the number of tags used and the number of bookmarks created was 0.56, although it was higher for users with bookmark collections smaller than 10 (0.74). For Flickr,[5] the correlation between distinct tags and photos was 0.518, and for del.icio.us,[4] no strong association existed between the number of bookmarks users had created and the number of tags they used in those bookmarks.

The high correlation for CiteULike suggests a strong linear relationship between the number of papers and the number of distinct tags for each user. This relationship could be due to the fact that as users tag more papers, the number of tags in their personal tag vocabulary increases.

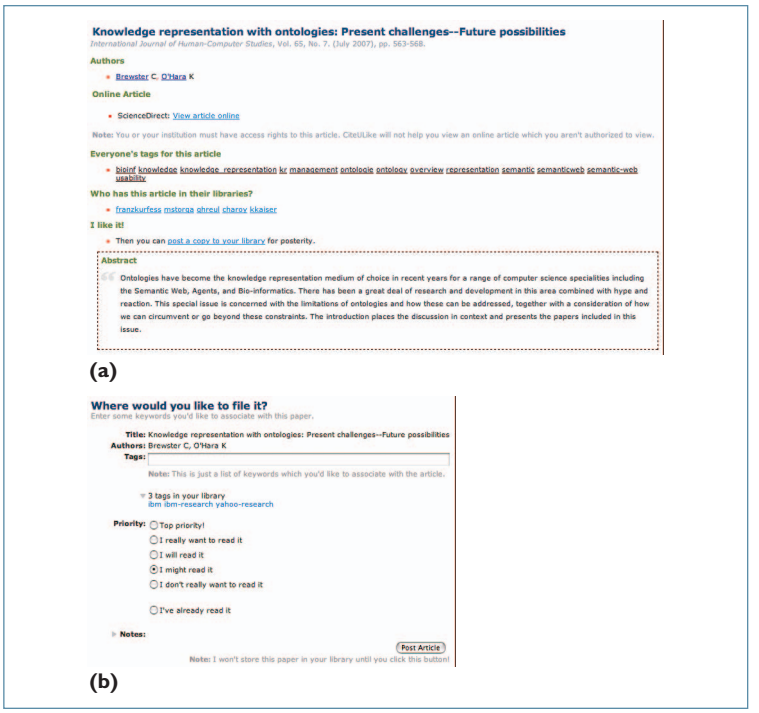

Figure 2. The CiteULike social bookmarking service. (a) A screenshot of the Web site shows a scholarly paper tagged in CiteULike. (b) The tagging page on CiteULike.

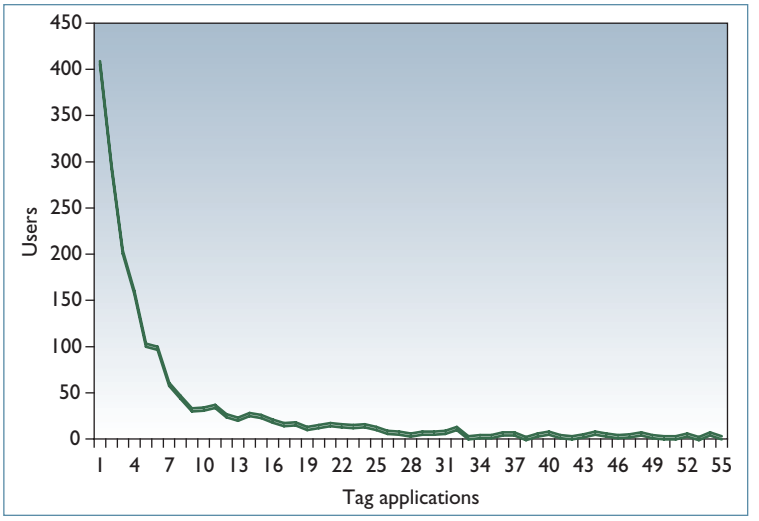

Figure 3. Number of users vs. number of tag applications. Relatively few users generated most of the tag applications.

Tag Growth

Social bookmarking services' premise is that users collaboratively generate and reuse tags. One way to index collaboration in social bookmarking services is to look at how users create new tags over time. We categorized the number of new tags per month, choosing months as the unit of temporal analysis. (A finer-grained denomination, such as days or weeks, would have resulted in too many data points to feasibly analyze visually.)

One form of tag vocabulary growth occurs at a diminishing rate over time,[5] which we can perhaps

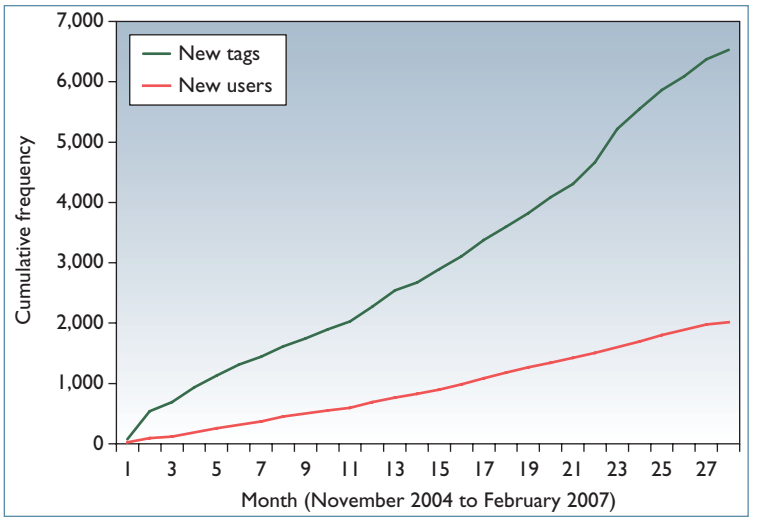

Figure 4. Cumulative frequency of new tags and new users over time. New tags and new users seem to be consistently growing in a linear fashion.

expect for a social bookmarking service, as it implies increasing stability in the tag vocabulary. However, for CiteULike, the tag vocabulary seems to be consistently growing. When we plotted the new tags' cumulative frequency (their aggregate summation) across time, the relationship was linear, as the green line in Figure 4 shows.

We think this consistent growth is due to the proportional increase in the number of new users. In the CiteULike data, we identified users as new when they applied a tag for the first time. We categorized new users across time (per month), and their cumulative frequency was a linear relationship (the red line in Figure 4), implying that they're also consistently growing over time.

To compare the cumulative frequencies of new tags and new users across time on the same scale, we calculated the cumulative frequency percentage. For new tags, we calculated the cumulative frequency of new tags per month as a percentage of the total number of tags for new users, and we calculated cumulative frequency of new users per month as a percentage of the total number of users.

The cumulative frequency percentages of new tags and new users over time are perfectly correlated (0.997), both growing at a linear rate and dependent on each other, which is consistent with our speculation that as new users apply tags, they create new ones.

Tag Reuse

For a social bookmarking service to be highly collaborative, we expect the tag vocabulary to converge and tag reuse to increase significantly over time. We can measure tag reuse in many ways — for example, a simple metric is to calculate the number of tag reuse applications:

tag reuse applications =
tag applications – distinct tags.

The minimum value for tag applications is the number of distinct tags, which implies that the minimum value for the number of tag reuse applications is zero (that is, there is no tag reuse). Using this metric, CiteULike had 25,715 tag reuse applications in our analysis.

This number doesn't tell us a whole lot about the amount of tag reuse, however. Thus, we use the more accurate and robust tag reuse metric Shilad Sen and colleagues developed for MovieLens,[3] one that calculates the number of users per tag according to the following formula:

tag reuse =
Σ (# of distinct users for each tag) / # of tags.

Given that each tag will have at least one associated user, the minimum value for tag reuse is 1.0 users per tag. For CiteULike, tag reuse was 1.59 users per tag. This is fairly low for tag reuse based on baseline figures from the MovieLens analysis.[3]

We also calculated how many tag reuse occurrences existed for each tag (number of tag applications per tag minus one). The average number of tag reuse occurrences was 3.9; however, the median and modal numbers were both zero. This indicates that most tags weren't reused, but a few tags were reused many times.

Figure 5a shows how many tags have been reused. The x-axis indicates tag reuse occurrences, whereas the y-axis indicates the number of tags. We've sorted the data in ascending order of tag reuse occurrences. For example, data point "A" indicates that 1,014 tags were reused once; data point "B" indicates that 514 tags were reused twice; and so on. The data resembles a power-law distribution: $y = 2043.6x^{-1.6727}$, $R^2 = 0.9469$ (the data set included 3,058 tags for a range of 1 to 48 tag reuse occurrences).

We also wanted to understand how many tags users were reusing from their personal collections (that is, how much a user reuses tags he or she has

applied before). The average number of tag reuse occurrences for each user was 8.5; the median and modal numbers were 5 and 1, respectively. This indicates that users were moderately reusing tags from their personal collections when tagging new papers. Figure 5b shows the results. Data point "A" indicates that 167 users reused one tag from their personal collections; data point "B" indicates that 136 users reused two tags from their personal collections; and so on. Again, the data resembled a power-law distribution: $y = 370.7x^{-1.3172}$, $R^2 = 0.8862$ (the data set included 879 users for a range of 1 to 49 tags reused).

Does CiteULike Support "Social" Bookmarking?

Although CiteULike supports tag reuse, many users didn't reuse tags from others' collections, although they reused tags from their own. We can explain this disparity at a human–computer interaction level. Clearly, the interface that CiteULike gives users during tagging affects their tagging behavior. When users tag papers, the interface lets them conveniently select and reuse tags from their personal collections; when they want to reuse tags from outside their collections, however, they can't view them during tagging.

As mentioned previously, the only way users can deliberately reuse tags from others' collection is to remember them from when they first viewed the article link; through mere coincidence, they might also reuse a tag. Thus, CiteULike doesn't explicitly support reuse through social transactions, which would explain why such tag reuse is low.

If social bookmarking services want to encourage greater tag reuse, they should pay particular attention to interface design. For example, in CiteSeer, we're now designing an integrated tagging interface such that users can see existing tags, from both their personal collections and others'.

Encouraging tag reuse requires not only an integrated tagging interface but also an appropriate tagging recommendation system. Not all existing tags are relevant to every paper; when the number of existing tags gets sufficiently large, users will be cognitively overloaded with respect to browsing and selecting relevant tags. Tag recommendation can address this problem by suggesting appropriate tags for papers based on several criteria. Currently, CiteULike presents the most frequently used tags (using visual enhancement – that is, a larger font size) in a user's personal collection to that user when he or she tags a paper. Although tag frequency is one heuristic for recommending tags, it doesn't have any bearing on those tags' relevance to the paper.

A more practical way to recommend tags is to compare similarities between papers and their associated tags. When a user is about to tag a new paper, an automatic tagging recommendation system can suggest relevant tags based on similarity measures between the new paper and existing ones. Note that different recommendation systems can also affect the amount of tag reuse.[3]

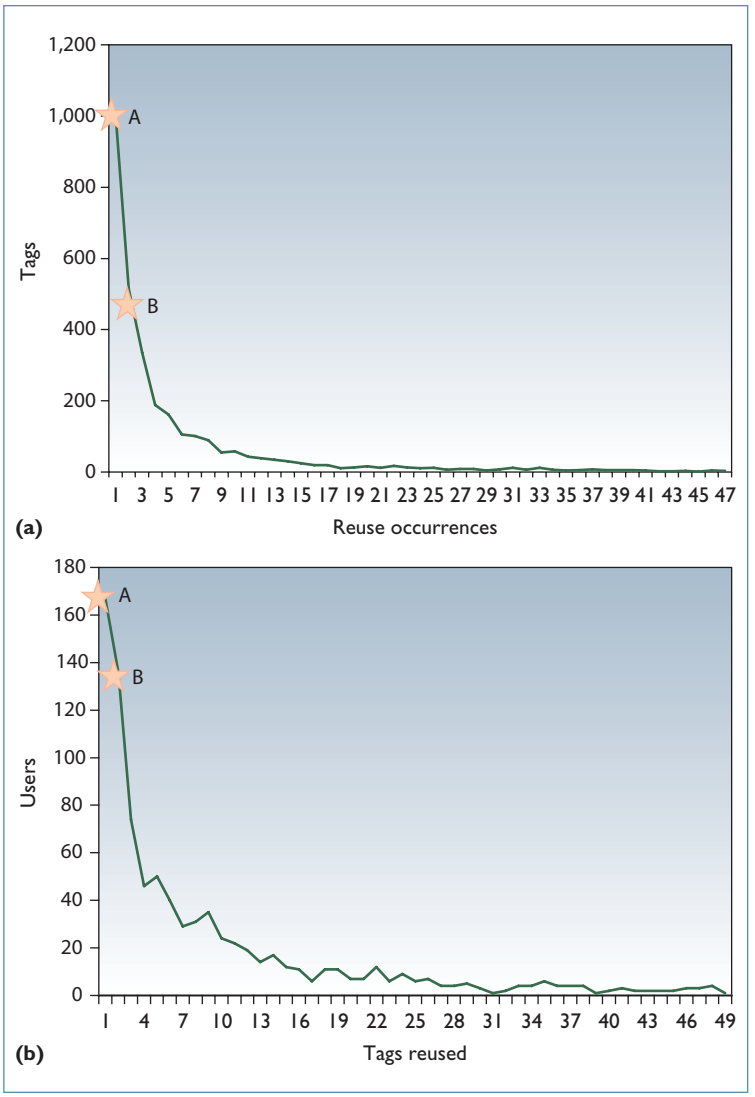

Figure 5. Tag reuse. (a) For tag reuse occurrences, "A" indicates that 1,014 tags were reused once; "B" indicates that 514 tags were reused twice. (b) Users and the frequency of reuse occurrences from their personal collections. "A" indicates that 167 users reused one tag; "B" indicates that 136 users reused two tags.

A Maturing Tag Vocabulary

Our data analysis showed that CiteULike's tag vocabulary is consistently growing at a linear rate. This might be one model for tag vocabulary growth, but if the goal is to have growth at a diminishing rate, a social bookmarking service must facilitate tag convergence.

An obvious way to do this is to facilitate tag reuse whereby users create fewer new tags and recycle more existing ones. It's plausible that if CiteULike had more tag reuse, its tag vocabulary would be converging.

On the other hand, CiteULike's tag vocabulary might still be maturing, which could also explain the lack of convergence. Although our data spans more than two years, CiteULike is a domain-specific social bookmarking service, thereby attracting a niche user group. Maturation time for CiteULike's vocabulary might be longer than other, general-interest social bookmarking services (such as del.icio.us) because the user population for niche services is much smaller; achieving critical mass for such services can take more time. The fact that CiteULike users are also growing at a linear rate provides evidence that the number of users and hence their tag vocabulary hasn't yet reached a relatively stable state.

Social Bookmarking Services and Digital Libraries

Related to the tag growth metric, tag reuse provides a direct interpretation of how often users recycle tags in a social bookmarking system. Both tag growth and tag reuse are important metrics for understanding how a tag vocabulary is evolving.

A social bookmarking system could have high tag growth but low tag reuse (as with CiteULike), low tag growth and low tag reuse (perhaps implying that the system isn't being used for tagging at all), low tag growth and high tag reuse (users are recycling previous tags and not creating new ones), and so on. Such an assessment lets social bookmarking service administrators gauge how people are using their system and allows designers to consider how to balance support for tag growth and tag reuse in their systems.

One way to think about and apply the tag growth and tag reuse metrics in concert is to adopt an activity-centric perspective. This perspective can help to balance the use of these metrics, depending on the user activities that the service is trying to support. For instance, if a social bookmarking service exists to support resource browsing based on users' growing interests over time, the tag growth metric is more important than tag reuse for ensuring that users are adding enough new tags to the system to maintain a critical mass based on their changing interests.

In our own research efforts toward developing a social bookmarking service for CiteSeer, we have started to adopt such an activity-centric perspective. Based on an initial requirements survey of CiteSeer users,[9] we determined that one primary user activity we want to support is the formation of social networks based on common tag usage. In this case, tag reuse is critical because we want to facilitate maximum tag sharing among users so that tag-based social networks are tightly knit and meaningful.

We can expect that user activity in social bookmarking services will vary with different domains. After all, CiteULike and del.icio.us users have different goals, are social and collaborative to different extents within their communities, assign varying importance to resources, and so on. Although we didn't compare user activity across such factors, we believe that tagging behavior between a domain-specific service such as CiteULike and other services has different characteristics. We think that domain-specific social bookmarking services need specialized research investigation to examine their pros and cons.

Our analysis of CiteULike in general has opened up several research paths for investigating social search through bookmarking services in scholarly digital libraries. Primarily, we think that we can further enhance social search through such services by supporting the formation of communities and subcommunities around tags and their associated papers. Facilitating online scientific collaboration among peers would be likely to contextualize tags in a more social setting. We're currently exploring such functionality with CiteSeer.

Acknowledgments

We thank Richard Cameron for providing us with the CiteULike data. The US National Science Foundation (CRI-0454052) supports the work presented in this article.

References

1. C.L. Giles et al., "CiteSeer: An Automatic Citation Index-

ing System," *Proc. Conf. Digital Libraries*, ACM Press, 1998, pp. 89–98.
2. C. Cattuto, "Semiotic Dynamics in Online Social Communities," *European Physical J. C*, vol. 46, 2006, pp. 33–37.
3. S. Sen et al., "Tagging Communities, Vocabulary, Evolution," *Proc. Conf. Computer Supported Cooperative Work*, ACM Press, 2006, pp. 181–190.
4. S.A. Golder and B.A. Huberman, "Usage Patterns of Collaborative Tagging Systems," *J. Information Science*, vol. 32, no. 2, 2006, pp. 198–208.
5. C. Marlow et al., "HT06, Tagging Paper, Taxonomy, Flickr, Academic Article, to Read," *Proc. Conf. Hypertext and Hypermedia*, ACM Press, 2006, pp. 31–40.
6. D.R. Millen et al., "Dogear: Social Bookmarking in the Enterprise," *Proc. Conf. Computer Human Interaction*, ACM Press, 2006, pp. 111–120.
7. L. Page et al., *The PageRank Citation Ranking: Bring Order to the Web*, tech report, Dept. of Computer Science, Stanford Univ., 1998.
8. S. Brin and L. Page, "The Anatomy of a Large-Scale Hypertextual Web Search Engine," *Proc. Conf. World Wide Web*, Elsevier, 1998, pp. 107–117.
9. U. Farooq et al., "Supporting Distributed Scientific Collaboration: Implications for Designing the CiteSeer Collaboratory," *Proc. Hawaii Int'l Conf. System Sciences*, IEEE CS Press, 2007, p. 26c.

Umer Farooq is a final-year PhD student in information sciences and technology at Pennsylvania State University. His dissertation research focuses on supporting creativity in computer-supported collaborative systems. Farooq has a BS in computer science from the National University of Computer and Emerging Sciences, Pakistan, and an MS in computer science from Virginia Tech. He is a student member of the ACM. Contact him at ufarooq@ist.psu.edu.

Yang Song is a fourth-year PhD student in computer science and engineering at Pennsylvania State University. His research interests include developing and applying statistical machine learning methods for text classification and clustering, pattern recognition, Web information retrieval, and data mining. Song has a BE in computer science from Zhejiang University, China. He is a student member of the IEEE. Contact him at yasong@cse.psu.edu.

John M. Carroll is the Edward M. Frymoyer Chair Professor of information sciences and technology at Pennsylvania State University. His research interests include methods and theory in human–computer interaction, particularly as applied to networking tools for collaborative learning and problem solving, and design of interactive information systems. Carroll has a PhD in experimental psychology from Columbia University. He is a fellow of the ACM, the IEEE, and the Human Factors and Ergonomics Society. Contact him at jcarroll@ist.psu.edu.

C. Lee Giles is the David Reese Professor of information sciences and technology at Pennsylvania State University. His research interests are in intelligent Web tools, search engines and information retrieval, digital libraries, Web services, knowledge and information extraction, and data mining. Giles has a PhD in optical sciences from the University of Arizona. He was a co-creator of the popular search tools, CiteSeer, for computer science, and ChemXSeer, for chemistry. He is a fellow of the ACM, the IEEE, and the International Neural Network Society. Contact him at giles@ist.psu.edu.