Understanding Research Trends in Conferences using PaperLens

Bongshin Lee [1,2] **Mary Czerwinski** [2] **George Robertson** [2] **Benjamin B. Bederson** [1]

[1]Human-Computer Interaction Lab
Computer Science Department,
University of Maryland,
College Park, MD 20742, USA
{bongshin, bederson}@cs.umd.edu
+1-301-405-7027

[2]Microsoft Research
One Microsoft Way
Redmond, WA 98052, USA
{marycz, ggr}@microsoft.com

ABSTRACT

PaperLens is a novel visualization that reveals trends, connections, and activity throughout a conference community. It tightly couples views across papers, authors, and references. PaperLens was developed to visualize 8 years (1995-2002) of InfoVis conference proceedings and was then extended to visualize 23 years (1982-2004) of the ACM SIGCHI conference proceedings. This paper describes how we designed PaperLens and analyzed the data from these two conferences, including some observations of patterns and relationships discovered using PaperLens. We also discuss the difficulties of handling incomplete, error-prone data. We then describe a user study we conducted to focus our redesign efforts along with the design changes we made to address usability issues. In addition, we summarize lessons learned in the process of design and scaling up to the larger set of CHI conference papers. The visualization contributes to the field by allowing users to discover research trends, patterns and relationships not possible with existing tools.

Author Keywords

Information visualization, Evaluation, Brushing, Timeline views, Piccolo.NET.

ACM Classification Keywords

H.2.8.c. Database Applications: Data and knowledge visualization; H.2.8.h. Database Applications: Interactive data exploration and discovery; H.5.2f. Information interfaces and presentation (HCI): Graphical user interfaces.

INTRODUCTION

Despite the rapid increase in the volume of scientific articles, the web makes them easily accessible. Furthermore, online scientific literature and digital libraries such as the ACM Digital Library (DL) [1], IEEE Xplore [8], CiteSeer [17], and the HCI Bibliography [7] provide broad bibliographical and full-text access to journals and conference proceedings. Users can search for a specific paper, typically by title or author. Some of the digital libraries may also be browsed by type of publication such as journals, transactions, and proceedings.

The ACM DL and CiteSeer show which papers are cited by a particular publication and which papers cite a particular publication. The ACM DL and the Digital Bibliography & Library Project (DBLP) [5] list all colleagues who have ever published with a particular author. Once users find a desired paper/author, they can easily access related papers/authors. However, it is often difficult to reconstruct navigation paths or to remember how a particular paper or author was found using these tools.

Envision [14] is a digital library augmented with a flexible user interface that provides a variety of visualization facilities, allowing users to explore patterns in the literature by making the context of the publications more apparent. IN-SPIRE's [20] Galaxy and ThemeView introduce visualizations of themes in document collections.

Most of the existing systems, however, are not designed to help users understand research trends. A few digital libraries provide some simple, statistical facts. For example, the HCI Bibliography shows the most frequently published authors and CiteSeer shows the most frequently cited papers/authors. However, the results are provided in the form of a long list, which only shows the name of authors or the title of the papers. Hence, simple analysis often requires extensive navigation and effort. For example, it is very difficult to find the author who has published the most in CHI, or to determine which papers referenced the most frequently cited author.

It is even more difficult to understand how researchers, topics, and outside research sources interact and influence research activity in general. Hence, Smeaton *et al*. [18] performed a content analysis of all papers published in the SIGIR proceedings to understand research trends and to

identify emerging and "hot" areas. Their focus was to determine what topic areas appear in the papers at the SIGIR conferences but not to visualize the results. They represented the clustering results in a table, whose rows and columns represent topics and years respectively. Each cell in the table contains the number of papers. The topics are sorted approximately in order of a combination of the year of their first appearance and the number of papers published. Since they color coded the cells by the number of papers, it is easy to recognize "hot" topics. However, users have to read through and compile the numbers to see the trends of topics. This makes it especially difficult to compare trends of several topics. They also did not include any citation analysis.

In practical terms, we are unable to answer interesting questions about our field of HCI with the current systems such as: Which topics have come and gone over the last 23 years of CHI? Which topics are currently hot in CHI? Which papers/authors are most frequently referenced by the CHI publications? How has the most frequently referenced papers/authors changed over time? What is the relationship between a given set of researchers?

Figure 1. PaperLens tightly couples views across papers, authors, and references: (a) Popularity of Topic (b) Selected Authors (c) Author List (d) Degrees of Separation Links (e) Paper List (f) Year by Year Top 10 Cited Papers/Authors.

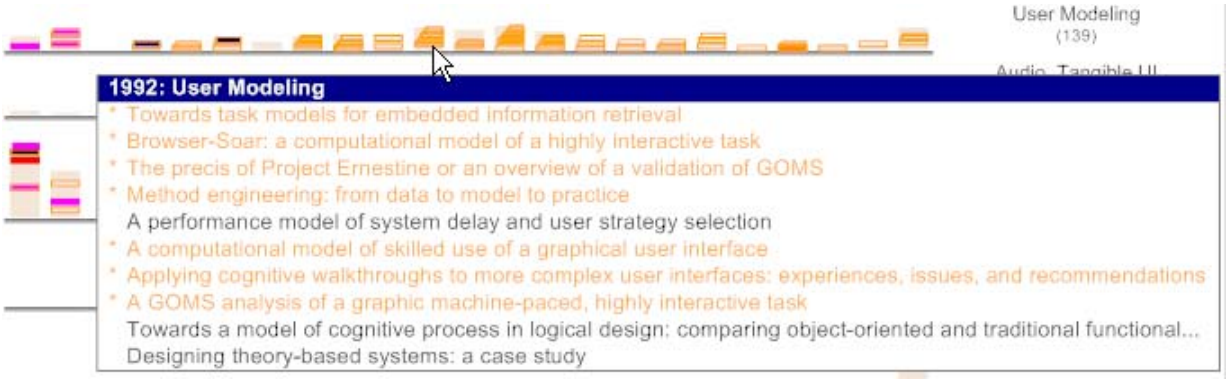

Figure 2. A pop-up menu showing the list of papers close to the current cursor position.

The IEEE InfoVis 2004 Conference chose to pose these kinds of questions about its history as the theme of the InfoVis 2004 Contest [9]. Three submissions to the contest [10,11,20] were given First Place awards. One of these was PaperLens, which we describe in this paper.

In order to address the questions, we developed a visualization called PaperLens, which allows researchers to see trends and topics in a field, in addition to influential papers and authors, all within a single screen visualization. PaperLens was developed to visualize 8 years (1995-2002) of InfoVis conference proceedings and then was extended to visualize 23 years (1982-2004) of ACM SIGCHI conference proceedings (see Figure 1).

In the following section, we describe the two datasets. Next we describe the basic design of PaperLens along with observations from both datasets. Then we describe a user study of the first version of PaperLens, followed by a description of design changes based on that user study. We conclude with a summary of lessons learned and future directions.

DATA ANALYSIS

Two Data Sets: InfoVis and CHI

The InfoVis 2004 contest [9] chairs provided the dataset containing metadata for 8 years (1995-2002) of InfoVis conference papers and their references. 315 authors published 155 papers in the InfoVis symposia. The metadata included author name(s), paper title, year of publication, and references for each paper. Some papers also had keywords, abstracts, references, and links to original papers.

The contest chairs collected all the available InfoVis publications and extracted their references by hand. They then found the referenced articles (if available) in the ACM Digital Library and collected the metadata for the referenced articles (if available) from the ACM Digital Library. Finally, they put everything together in one XML file. After the contest chairs released the dataset, other researchers helped them clean up the data set. For example, a duplicate authors' map was provided.

Two well-known sources of public citation information, CiteSeer [17] and ParaCite [15], automatically extract references from PDF files through complex heuristics implemented as Perl scripts. However, these scripts are unreliable and fail to work on PDF files containing bitmaps, such as the proceedings of InfoVis from 1995 to 1997. Therefore, the references had to be extracted by hand.

Once we visualized the InfoVis data, ACM kindly provided the dataset containing metadata for 23 years (1982-2004) of CHI papers. Though the data for the papers published in 1984 is missing, the remaining dataset included not only full papers but also short papers, demos, and videos. The complete dataset includes 6309 authors and 4073 papers. The reference data was problematic, and only 43% of the references had a paper identifier assigned by the ACM DL. While we did have the complete reference text, we chose to focus on the visualization, and therefore we did not make a further effort to parse or otherwise improve the reference data. However, we did write a simple Perl script to retrieve the necessary metadata such as paper source, year of publication, title, and authors from the ACM DL. We also had to manually clean up the duplicate author names (e.g., Stuart Card in addition to S.K. Card, etc.).

Topic Clustering

We used standard, internally developed topic clustering technology originally developed for site administrators to help build and maintain category hierarchies. The text-clustering component suggests a hierarchically organized set of categories when no explicit structure exists. We used titles, references, and keywords in the clustering process. For InfoVis, we weighed the titles more heavily to get a better clustering result. A standard list of stop words, months of the year, journal and proceeding titles, and version and page numbers were removed from influencing the cluster results.

Five InfoVis and 22 CHI clusters emerged from using the clustering tool. We used PaperLens in the process of manually naming each cluster by investigating papers and authors in the cluster. For CHI data, some topics were

divided into several clusters, which we combined into one cluster, but, we did not move individual papers into other clusters. This resulted in some papers being placed in odd clusters but is typical of any clustering solution. We ended up with 15 CHI clusters summarized in Table 1 and sorted by the number of papers in each.

Clusters	Number of Papers
Lab Reports, Applications, Web	618
Multimodal UI	601
CSCW	512
Miscellaneous	407
Usability	334
InfoVis	323
Cognitive Factors in Design	241
Anthropomorphism	209
End User Programming	160
Target Acquisition	140
User Modeling	139
Audio, Tangible UI	130
User Centered Design	119
VR, Input Devices	75
UIMS	65

Table 1. 15 Clusters of CHI papers by topic.

PAPERLENS INTERFACE

The goal of PaperLens is for the novice or expert to gain some insights as to how a field's topics and research activities have changed over time. PaperLens (Figure 1) consists of 6 main parts: a) popularity of topic; b) selected authors; c) author list; d) paper list; e) degrees of separation links; and f) year by year top 10 cited papers/authors. In this section, we describe the PaperLens user interface along with the interesting patterns and relationships we discovered.

Evolution of Topics

In the *popularity of topic* view (Figure 1a), we organized papers by their topic and year. The light beige bars represent a group of papers whose height is proportional to the number of papers in the group. This allows an interested user to quickly see trends of the topics over time. As can be seen from Figure 1a describing the CHI dataset, the InfoVis category (10th from the top) emerged in the late 1980's and then stayed steady from the early 1990's. The topics of CSCW and Anthropomorphism exhibited steady increases in popularity while the UIMS category almost died out around 1995 (11th through 13th from the top, respectively).

Furthermore, by hovering on an individual topic title, the years when that topic was most popular are signified by a brown border around the relevant column in the *popularity of topic* view (Figure 3). To help users see when the topic was popular, we show the year above the bar. For example, the Cognitive Factors in Design category was the most popular in early years (Figure 3a), Lab Reports etc. in the middle years (Figure 3b), and CSCW and Multimodal UI became more popular in recent years (Figure 3c, Figure 3d).

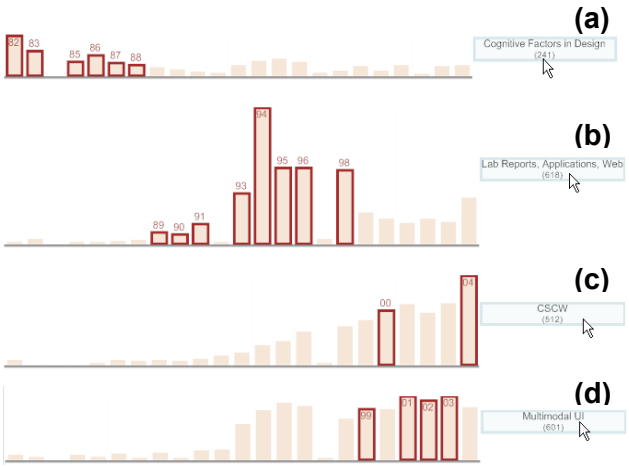

Figure 3. Highlighting of the most common topics: (a) Cognitive Factors in Design (b) Lab Reports, Application, Web (c) CSCW (d) Multimodal UI.

Easy Access to Papers/Authors

PaperLens provides a way to search for specific papers – a simple substring match by title. By typing in a keyword, such as "3d", the entire visualization is filtered to only show information related to papers that match the search string in their titles. PaperLens also enables users to get a list of papers by topic or by authors. By selecting a topic in the display, the list of all papers in that area is shown in the *paper list* (Figure 1e). We also show the number of citations for each paper in the *paper list* to show the influence of the paper.

When the user selects authors from the authors list, they are shown in the *selected authors* area (Figure 1b). The author name search is a bit different than the paper search in that its substring search only works from the initial letter of a first or last name. The current search hit is signified by a pink background and users can iterate through multiple hits by clicking the "Find Next" button.

Once authors are added to the *selected authors* area, all of the papers by them are shown in the *paper list* and are highlighted in the *popularity of topic* view, matched to the author by color coding. The color coding enables users to see which topic area a particular author fits in. For example, Stuart Card has mainly published in the InfoVis area, as seen by his representative red color coding seen within the *popularity of topic* view (Figure 1). We used black when two or more selected authors wrote a paper together. By selecting two authors, we could immediately see whether they had ever published together.

For the InfoVis proceedings data, we were able to devote one rectangle to any individual paper in the *popularity of topic* view. If we stack the small rectangles representing papers without overlap for the CHI proceedings, however, the height for each paper is less than 1 pixel, which turned out to be very difficult to recognize. So, we decided to render each rectangle 4 pixels high, and to raise highlighted rectangles to the front. Even so, when several papers are highlighted, the corresponding rectangles sometimes overlap. So when overlapped, they are shifted one pixel to the right (Figure 2). Since overlapping made it difficult to select a paper from the *popularity of topic* view for the CHI data, we provided a pop-up list menu showing the papers close to the current cursor position. Paper titles are matched to the highlighted paper by color coding. The pop-up menu appears upon mouse-over like a tool tip and is pinned down when the user clicks the *popularity of topic* view.

Once the menu is pinned down, it works just like a popup menu. The user can select a paper by single click from the list. When a paper is selected, its authors are automatically picked from the *author list* (Figure 1c) and added to the *selected authors* area. In addition, we highlight papers that have referenced the selected paper via orange highlighting in the *popularity of topic* area. A double-click takes the user to the ACM Portal with a link to the paper. (For the InfoVis proceedings, accessing a paper was simply a double-click on the paper's rectangle; the title and author name(s) were provided with a tool tip on mouse over as shown in Figure 10).

Most Frequently Published Authors

We show the number of papers published by an author in the *author list* (Figure 1c). Users can sort the *author list* by the number of papers and immediately see who has published the most. For example, the most prolific author is Brad Myers who has contributed 41 papers to the CHI conference. When the user selects a topic in the *popularity of topic* view, a column having the number of papers in that topic is added to the *author list*. Now users can see who has published the most on each topic. For example, Jonathan Grudin was ranked #1 with 16 papers for the CSCW topic. 16 papers out of 26 were published by Jonathan within the CSCW topic category.

Relationships between Authors

A co-author collaboration graph is often used to find the relationship between individual authors and the center of the community, i.e., the author that has the shortest average path length to all other authors in the graph [4,13,18]. The graph among InfoVis authors, however, is too fragmented to give any useful insights. For the InfoVis data, S. F. Roth, the center of the graph, has published 5 papers with 13 co-authors and has only 19 related colleagues among 315 possible individuals. Even though the largest component of the collaboration graph for CHI authors is bigger than that for InfoVis authors, it is still fragmented with many small components. We therefore decided to display all of the related colleagues for an author when she is selected by the user. We computed the shortest path length between two authors on demand and called it *degrees of separation links* (Figure 1d and Figure 4).

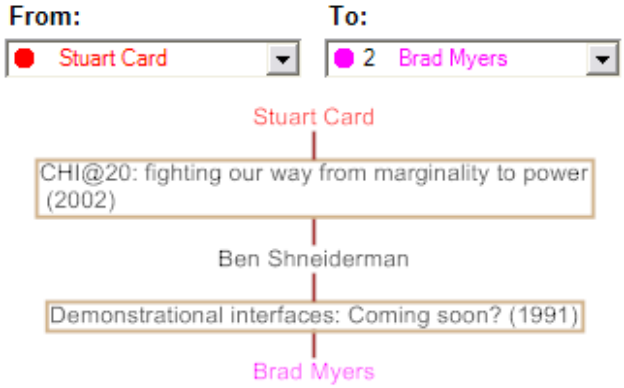

Figure 4. *Degrees of Separation Links* shows how Stuart Card and Brad Myers are connected by a co-author relationship.

The *degrees of separation links* view plays an important role in showing the relationships between authors in the CHI community. The *From* combo-box contains the same list of authors as the *selected authors* area. Once an author is selected from the *From* combo-box, we display all the selected author's related colleagues in the *To* combo-box with the corresponding degrees of separation. When an author is selected from the *To* combo-box, we display the shortest path between two authors in our dataset through their *degrees of separation links*. For example, Stuart Card and Brad Myers are connected indirectly to each other because they have each co-authored a paper with Ben Shneiderman

Most Frequently Referenced Papers

One of the interesting questions we wanted to answer was "Which papers/authors are most often referenced?" because this is one important metric indicating influential papers/authors. In addition to counting the number of references overall, we computed them by year and by topic to show trends.

The top 10 most frequently cited papers are viewable in the *overall* list within the *year by year top 10 cited papers* view. The papers included here are all the papers available to us in this study, including all the CHI papers and all the papers that the CHI papers reference. One can also see the detailed information for each of the top 10 papers and how it was referenced over time by hovering on an individual paper. It is easy to see that the most frequently cited publication by CHI authors to date is "The Psychology of Human Computer Interaction" by Card *et al.*, which was published in the Journal of Electronic Materials (Figure 5a). Selecting a paper from

the *year by year top 10 cited papers* view (Figure 1f) has same effect as selecting a paper for the popularity of topic view.

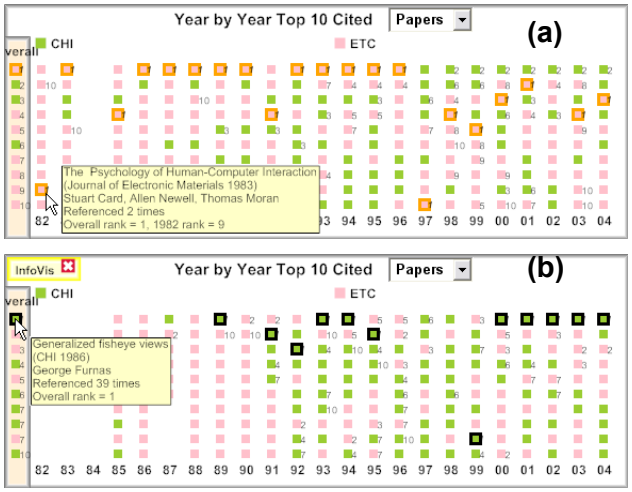

Figure 5. (a) The Psychology of HCI paper is the most referenced paper by all CHI papers, (b) Fisheye Views papers is the most referenced by InfoVis CHI papers

We distinguished the important CHI papers with the color green. It can immediately be seen that CHI publications themselves have been a significant source of inspiration for the CHI community to date and that 3 of the top 10 papers most frequently cited are CHI publications. Most years do have at least one CHI paper in the top 10, with the notable exception of 1982, which was CHI's first year.

When the user selects a topic from the *popularity of topic* view, the *year by year top 10 citations* area is filtered to only show the frequent citations for that topic area. In this way, PaperLens allows the user to quickly discover the most influential papers in a particular topic area. For instance, for the InfoVis topic, the Generalized Fisheye Views paper written by George Furnas was the #1 most frequently cited paper (Figure 5b).

Most Frequently Referenced Authors and Their Papers

Ranking the frequent citations by author shows frequently cited authors that either have or have not published in CHI. Since we colored the authors pink if they have not ever published in CHI, it can be seen that all overall top 10 cited authors have published in CHI, even for the End User Programming topic (Figure 6). However, for several years, many top 10 frequently cited authors for the End User Programming have not published in CHI (Figure 6b). Selecting an author from the *year by year top 10 cited authors* view (Figure 1f) shows not only any papers selected authors have published in CHI, but also those papers that have referenced them via orange highlighting in the *popularity of topic* area (Figure 1). The user can immediately discover which areas were most influenced by the selected author. For instance, for End User Programming, Brad Myers was the most frequently cited author (Figure 6b).

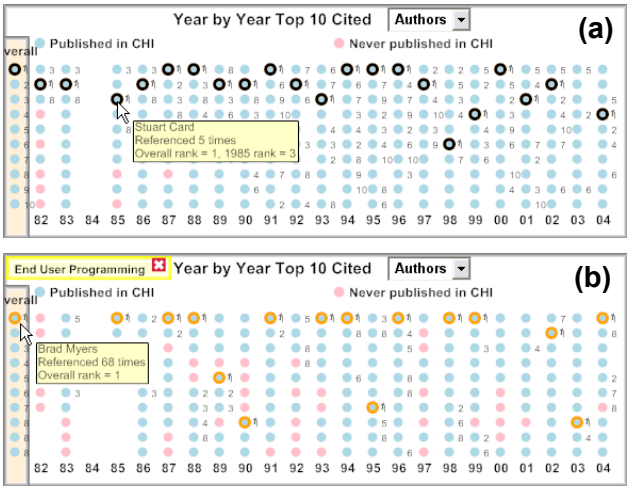

Figure 6. (a) Stuart Card is the most frequently cited author by all CHI papers and (b) Brad Myers is the most frequently cited author for End User Programming

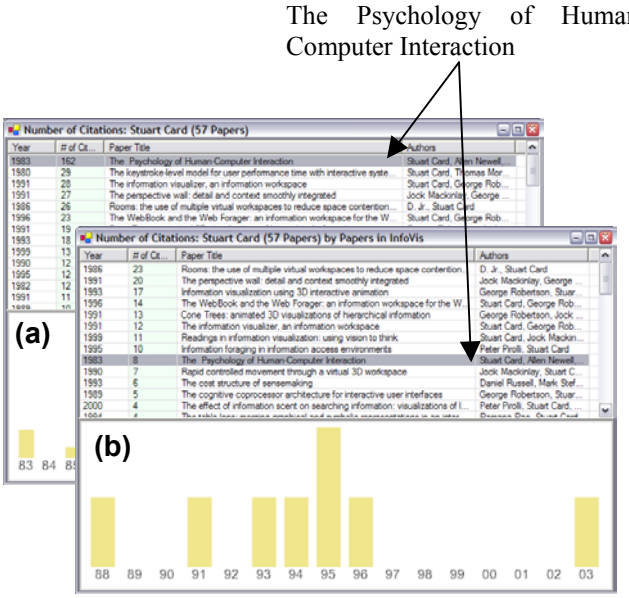

Figure 7. Number of citations of the papers written by Stuart Card (a) by all CHI publications and (b) by the papers in the InfoVis topic. "The Psychology of Human Computer Interaction" was referenced 162 times by all CHI publications but only 8 times by the papers in the InfoVis topic.

Once we know who the most referenced authors are, it would be interesting to see how many times each of their papers is referenced. We decided to add a view similar to that available in CiteSeer. A double-click on an author in the *year by year top 10 cited authors* view opens a *number of citations* view (Figure 7), which shows the number of citations of the papers written by the selected author. Furthermore, when a topic is selected by the user,

the number of citations is counted only by the papers in that topic area. For example, Stuart Card was the most frequently cited author by all CHI authors and "The Psychology of Human Computer Interaction" was the seminal contribution (Figure 7a). While he was also the most frequently cited author for the InfoVis topic, that paper was referenced only 8 times and other papers such as the Perspective Wall and Cone Trees papers were referenced more often (Figure 7b). It also shows the distribution of references by year at the bottom half of the window. This view has advantages over CiteSeer in that it is organized by author instead of by an author's individual paper.

Implementation

PaperLens was implemented in C# and runs on any standard Windows PC. In addition to the main data file in an XML format, we used an application-specific version of the datasets in simple tab-separated text files. We loaded the entire dataset into memory. All the graphical views are implemented with Piccolo.NET, a shared source toolkit that supports scalable structured 2D graphics [2,16].

Weaknesses

As is often the case with powerful and new visualization tools, PaperLens requires some learning time. New users need to learn how to decode the various color mappings and highlighting. They are also required to understand how views are coupled because all views and interactions are symmetrical and can be used for both input and output.

USER STUDY

In order to understand how useful our original user interface design ideas were, and to help us iterate to a more usable design, a user study was carried out using the InfoVis dataset and the first iteration prototype. Eight researchers (including 1 pilot subject) who were on the "fringe" of the information visualization community but had never published at the InfoVis conference itself were recruited internally. Four of the researchers were computer science graduate student interns, and four were full time researchers, and all were interested and actively working in the area of HCI. Ages of the participants ranged from 24 to 42. The pilot data is included in the discussion of the usability issues observed, but not in the reporting of the experimental task data, as the task set was altered slightly after the pilot.

Participants were given a brief tutorial of the system, spending no longer than 20 minutes reading about and interacting with features of the system. This segment of the study was considered "think aloud", and any usability issues they experienced during this walk through of the system were noted by the experimenter.

Next, the participants were asked to carry out several experimental tasks with the system, which were timed and scored for correctness. The tasks were meant to not only evaluate the usefulness and usability of many features of the initial prototype, but also to determine areas that might need to be redesigned or even eliminated as we attempted to scale up to the much larger CHI proceedings dataset. All users carried the tasks out sequentially, as quickly as they were able. Once a task was over, participants were allowed to discuss what did or did not work well with the system in terms of efficiently completing the task, and the experimenter recorded these comments. Once all of the tasks were completed, users were asked to fill out a satisfaction questionnaire about PaperLens. All sessions were completed with participants run individually and lasted no more than one hour. Participants were provided with a coupon for a free lunch or dinner from the cafeteria for their participation. The list of the tasks follows.

1) Who published the only paper on Graph Visualization in 1998?
2) How many papers did S. K. Card publish at InfoVis over the 8 years in our database?
3) Who were George Robertson's coauthors on his only paper in the database?
4) How many degrees of separation exist between S. F. Roth and S. G. Eick?
5) Which topic area has enjoyed gradual growth over the last 8 years?
6) Which topic area has all but died out in terms of papers published on that topic over the last 8 years?
7) Which topic area has had many more papers published on that topic during the last 2 years in our database?
8) Which authors are in the top 10 most frequently cited list but have not published at InfoVis?
9) How many papers of the top 10 most frequently cited papers are from InfoVis?
10) How many papers in the top 10 most frequently cited list are from CHI?
11) Which topic area references the most frequently cited paper most often?
12) Go to the most frequently cited InfoVis paper and read it's abstract.
13) In the Dynamic Queries topic area, which author is the most frequently cited?
14) What was the last year that S. K. Card published in this database?
15) Who was the most frequently cited author in 2001?
16) How many papers did J. Mackinlay and S. K. Card publish together at InfoVis over the 8 years in our database?

Results

Task Times and Errors

Overall, participants were able to correctly answer the tasks used in the study 97% of the time. There were only 5 incorrect answers provided out of a possible 112 questions across participants. Three participants each gave one wrong answer and one participant incorrectly

answered 2 questions. Each incorrect answer was from a different question for each participant. Incorrect answer times were not included in the task time analysis.

Overall, average task times were fast, with only the last task taking much longer than the others (1 minute and 5 seconds, on average). This task required users to figure out how many papers J. Mackinlay and S. K. Card published together at InfoVis, which required users to remember that black color coding was used to signify multiple co-authors on a paper. Most other tasks were performed in less than 20 seconds, as can been seen from Figure 8. Usability issues leading to longer task times will be discussed next.

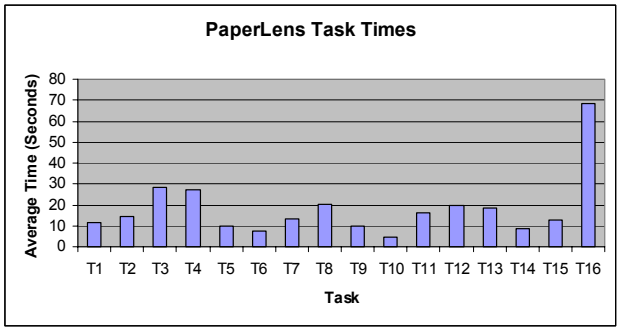

Figure 8. Average task times using PaperLens with the InfoVis proceedings.

Usability Issues

Several usability issues were observed that needed to be addressed through design iteration. These issues were prioritized based on how many of the participants encountered them and the severity of the issue in terms of being able to easily recover from it, or based on how long the issue delayed finding an answer to a task. The highest priority issues centered on the way searching for authors was implemented: in this prototype it was initially a string-based search that did not allow the user to search for first or last names separately, and found substring matches anywhere in the name (not just at the beginning of names). We addressed this issue by providing columns for searching on the first and last name, in addition to fixing the way our substring matches worked (from the beginning of names).

There were several high priority issues observed where our system did not behave "symmetrically". In other words, if you could launch a paper from one list view, you should be able to open it from any list view, etc. All of these symmetry issues have been addressed in the redesigned system.

Finally, users gave us feedback concerning the *degrees of separation list and links* views (Figure 9) — while some users liked the links view, others thought that it was more "recreational". There were also some issues with the default way the degrees of separation were being displayed in the links view (e.g., picking the longest link chain by default) that had to be addressed. To help alleviate usability issues in this area, in addition to freeing up screen real estate, we combined the list and links views into one view (now called *degrees of separation links* in the current system) and allowed the user to pick the degrees of separation between any selected author and related people.

Other design changes were made to fix less severe usability problems. While there were some ways to clear selections for each view, they were not consistent. Furthermore, there was no "home" button that cleared all selections. Many participants also wanted to toggle the selections for the topic label, year, paper, and author. We addressed these issues by adding a "home" button and making all selections "toggle-able". One more issue about search was the desire to cycle through the multiple matches. Iterating through search hits using the "Enter" key turned out not to be intuitive. We instead added "Find Next" button and used "Enter" for selection.

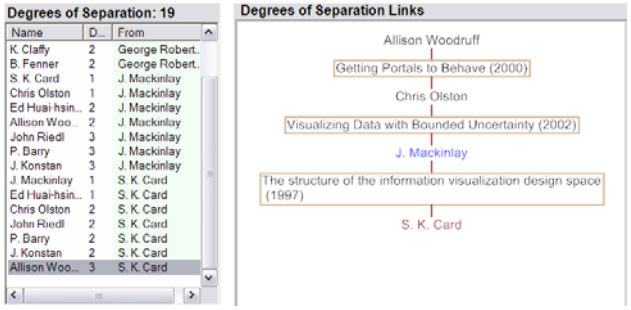

Figure 9. *Degrees of Separation List and Links* views from the earlier prototype.

User Satisfaction Ratings

The satisfaction questionnaire that users filled out at the end of the study session was analyzed. All ratings were on a 1-7 Likert scale, with 1=Disagree and 7=Agree. The average ratings are shown in Table 3 below. Overall, the system was rated fairly highly for a first time iteration of user testing. However, there was clear user frustration around the ease of learning the system, its look and organization, the *degrees of separation* area, and error recovery. In other words, the usability issues that were identified were fairly well corroborated in the satisfaction ratings. It is hoped that the redesign changes will go a long way to alleviate these satisfaction problems.

Overall, I am satisfied with this system.	5.3
It was simple to use this system.	4.3
I was able to efficiently complete the tasks and scenarios using this system.	6
I felt comfortable using this system.	5.3
It was easy to learn to use this system.	4.4
The "look" of this system was pleasant.	4.9

I liked using this system.	5.9
The organization of information in this system was clear.	3.9
Whenever I made a mistake using this system, I could recover quickly and easily.	4.6
It was easy to discover trends of the topics using the "Popularity of Topic" view.	6.3
It was easy to discover relationships between authors using the "Degrees of Separation Links" view.	4.1
It was easy to discover the most referenced papers/authors using the "Year by Year Top 10 Cited Papers/Authors" view.	6.9
Overall, highlighting on the screen was helpful.	6.1
Overall, the use of color was appropriate.	5.3

Table 3. Average Likert scale ratings for PaperLens, using the scale of 1=Disagree, 7=Agree.

LESSONS LEARNED

There are two core things we learned in the design and use of PaperLens. The first is that sometimes simple is good. Our first thoughts when confronted with the problems described in the InfoVis contest were to build some kind of graph visualization tool such as Gem-3D [3], dot [6], H3 [12], or NicheWorks [19] that showed all the data and all the relationships at once. But we suspected viewing too much information at once could be overwhelming and that the approach would quickly become unwieldy. Furthermore, there is no efficient way to show the trends of topics with graph visualizations.

We instead opted for a simpler design with an abstract overview of the full dataset but not with all the relationships visible. We also oriented the design around several small and simple tightly coupled views which, together, provide the user with powerful capabilities. While these design ideas have certainly appeared before, PaperLens brings them together in a unique fashion to address an important problem of concern to HCI researchers. To summarize, we think the key elements of the PaperLens design that make it work well are:

- An abstract overview
- Multiple small and simple components to best show the different aspects of the data
- Relationships shown through interactivity and tightly coupled components
- All visual elements are laid out along axes with well defined metrics

The second thing we learned is about issues in scaling up the visualization. For the InfoVis data, which has only 155 papers, we could use a square to represent each paper. This enabled the user to select a paper by a single click and was viewed positively by many participants. We also used a fisheye technique (Figure 10) to help people reveal the individual paper titles for that year by topic when the user clicked on a year. However, when we tried a similar approach (a rectangle instead of a square) for the CHI data, the height of the rectangle was too small (less than one pixel) to recognize and select it as a target. The fisheye technique did not work either because the number of papers for which we could show titles in one column was less than 70, and we needed to be able to show as many as 150.

Figure 10. For the earlier prototype, a fisheye technique is used to help people reveal the individual paper titles for that year by topic when the user clicked on a year.

Some users were concerned that there were so many different colors in the original user interface. We used 8 different colors to represent authors for the InfoVis data because 8 is the maximum number of authors of one paper. For the CHI data, the maximum number of authors is 15. We suspected that it would not be useful to have 15 different colors to distinguish authors from each other. We decided to use a single color if the number of selected authors is larger than or equal to 5.

FUTURE WORK

As future work, it would be interesting to allow a user to display any category of publication within an electronic library like the ACM Digital Library. Given the difficulty faced simply scaling up from InfoVis to CHI, it would be an excellent research exercise to move toward another order of magnitude in terms of the number of documents, authors and references that would require. At that point, the tool could be used more as a portal to move in and out of different classes or document types, journals or

conferences. We are planning to examine ways to scale the visualization to a much larger dataset of documents such as ACM DL with many other kinds of metadata.

In addition, it would be interesting to explore showing richer relationships among more than two authors, in other words, co-authorship graph visualizations might be useful to explore as the networks get larger, but they must be designed and demonstrated to be usable.

CONCLUSION

PaperLens is a visualization tool that allows users to see trends and topics in a field, in addition to influential papers and authors. We described two design iterations, the first designed for a smaller set of conference papers (InfoVis) and the second designed for a larger set of papers (CHI). The design iteration was driven both by user studies and by dealing with issues of scale. We found that PaperLens' design, which shows relationships within a complex network through interactive highlighting, is a compelling alternative to a more common network visualization with a node-link diagram.

Using PaperLens, we have discovered many interesting patterns and relationships which could not have been revealed using existing tools. For example, CHI was the most influential source of references used by InfoVis authors to date. George Robertson and George Furnas, both influential in the InfoVis proceedings, have only published once or never, respectively, at InfoVis. In the CHI proceedings, we see many similar trends and patterns across the field of HCI, as discussed above.

PaperLens' power comes from tightly coupled views across papers, authors, and references. The ability to query by time and topic has enabled novel views that we found very beneficial in our explorations of this domain.

ACKNOWLEDGMENTS

We would like to thank the InfoVis 2004 contest chairs, Jean-Daniel Fekete, Catherine Plaisant, and Georges Grinstein, for providing the contest data and ACM for providing the SIGCHI proceeding data. We also would like to thank Aaron Clamage who ported Piccolo from Java to .NET, and the participants of our user studies.

REFERENCES

1. ACM Digital Library. http://portal.acm.org
2. Bederson, B.B., Grosjean, J., and Meyer, J. (2004) Toolkit Design for Interactive Structured Graphics. *IEEE Transactions on Software Engineering*, Vol 30, No. 8, 535-546.
3. Bruß, I. and Frick, A. (1995) Fast Interactive 3-D Graph Visualization. In *Proceedings of the Symposium on Graph Drawing (GD '95)*, Lecture Notes in Computer Science 1027, 99–110. Springer-Verlag,
4. Chen, C. and Carr, L. (1999) Trailblazing the Literature of Hypertext: Author Co-Citation Analysis. In *Proceedings of the 10th ACM Conference on Hypertext and hypermedia*, 51-60.
5. DBLP Computer Science Bibliography. http://dblp.uni-trier.de
6. Gansner, E.R., Koutsofois, E., North, S.C. and Vo, K.-P. (1993) A Technique for Drawing Directed Graphs. *IEEE Transactions on Software Engineering*, Vol. 19 No. 3, 214–229.
7. HCI Bibliography. http://www.hcibib.org
8. IEEE Xplore. http://ieeexplore.ieee.org
9. InfoVis 2004 Contest: The History of InfoVis http://www.cs.umd.edu/hcil/iv04contest.
10. Ke, W., Borner, K., and Viswanath, L. (2004), Major Information Visualization Authors, Papers, and Topics in the ACM Library. In *Poster Compendium of InfoVis 2004.*
11. Lee, B., Czerwinski, M., Robertson, G., and Bederson, B.B. (2004) Understanding Eight Years of InfoVis Conferences Using PaperLens. In *Poster Compendium of InfoVis 2004.*
12. Munzner, T. (1997) H3: Laying Out Large Directed Graphs in 3D Hyperbolic Space. In *Proceedings of InfoVis 1997*, 2-10.
13. Nascimento, M.A., Sander, J., and Pound, J. (2003) Analysis of SIGMOD's CoAuthorship Graph. *SIGMOD Record*, 32(3).
14. Nowell, L.T., France, R.K., and Hix, D. (1997) Exploring Search Results with Envision. *Ext. Abstracts CHI 1997*, ACM Press, 14-15.
15. ParaCite. http://paracite.eprints.org/
16. Piccolo.NET. http://www.cs.umd.edu/hcil/piccolo
17. Scientific Literature Digital Library. http://citeseer.ist.psu.edu/cs
18. Smeaton, A.F., Keogh, G., Gurrin, C., McDonald, K., and Sødring, T. (2003) Analysis of papers from twenty-five years of SIGIR conferences: what have we been doing for the last quarter of a century?. *ACM SIGIR Forum*, 49-53.
19. Wills, G. J. (1997) NicheWorks – Interactive Visualization of Very Large Graphs. In *Proceedings of the Symposium on Graph Drawing (GD '97)*, Lecture Notes in Computer Science 1353, 403–414. Springer-Verlag.
20. Wong, P.C., Hetzler, B., Posse, C., Whiting, M., Havre, S., Cramer, N., Shah, A., Singhal, M., Turner, A., and Thomas, J. (2004) IN-SPIRE InfoVis 2004 Contest Entry. In *Poster Compendium of InfoVis 2004.*