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

[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 CHI conference proceedings. This paper describes how we analyzed the data and designed PaperLens. We also describe a user study to focus our redesign efforts along with the design changes we made to address usability issues. We summarize lessons learned in the process of design and scaling up to the larger set of CHI conference papers.

Author Keywords

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

ACM Classification Keywords

H.5.2 User Interfaces---Graphical user interfaces (GUI), H.5.2 User Interfaces---Evaluation/methodology, H.2.8. Database Applications---Data mining.

INTRODUCTION

Online digital libraries such as the ACM Digital Library (DL) [1] provide broad bibliographical and full-text access to journals and conference proceedings. The ACM DL shows which papers cite or are cited by a particular publication. It also lists all colleagues who have ever published with a particular author. This enables users to access related papers/authors once they find a desired paper/author. However, it is often difficult to reconstruct navigation paths and to remember how a particular paper/author was found using these tools.

Envision [6] 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. Galaxy and ThemeView introduce visualizations of themes in document collections [9]. Most existing systems, however, are not designed to help users understand research trends. A few digital libraries provide some simple, statistical facts such as the most frequently cited papers/authors. However, simple analysis often requires extensive navigation and effort since the results are provided in the form of a long list.

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*. [8] performed a content analysis of papers published in SIGIR proceedings to understand research trends. Their focus was to determine what topic areas appear but not to visualize the results. They also did not include any citation analysis.

In practical terms, we are unable to answer interesting questions with the current systems such as: Which topics have come and gone over the last 23 years of CHI? What is the relationship between a given set of researchers? The IEEE InfoVis 2004 Conference chose to pose these kinds of questions about its history as the theme of the InfoVis 2004 Contest [3]. 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.

DATA ANALYSIS

The InfoVis 2004 contest chairs provided a dataset containing metadata for 8 years of InfoVis conference papers and references. They collected all the available InfoVis publications and extracted their references by hand. They found the referenced articles and metadata (if available) in the ACM DL. Finally, they put everything together in one XML file. After the contest chairs released the dataset, other researchers helped them clean up the data.

Once we visualized the InfoVis data, ACM kindly provided the dataset containing metadata for 23 years of CHI papers. The dataset included full papers, short papers, demos, and videos. The reference data was problematic, and only 43% of the references had a paper identifier assigned. While we had the complete reference text, we focused on the visualization, and did not undertake further effort to improve reference data. However, we did write a simple Perl script to retrieve the necessary metadata such as paper source, year of publication, and authors from the ACM DL.

CHI 2005, April 2–7, 2005, Portland, Oregon, USA.
ACM 1-59593-002-7/05/0004.

To identify research topics, we used standard, internally developed topic clustering technology. The statistical model underlying the code is called a mixture model [5]. The technology was originally developed for site administrators to help build and maintain category hierarchies. The text-clustering component suggests a set of categories when no explicit structure exists. We used titles, references, and keywords in the clustering process. 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 the 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 the 15 CHI clusters shown in Figure 1.

PAPERLENS INTERFACE

In the *popularity of topic* view (Figure 1a) we organized papers by their topic and year. Hence, we can easily capture trends in the topics. For example, the InfoVis category (10th from the top) emerged in the late 1980's and then has remained steady in terms of publications from the early 1990's.

PaperLens enables users to get a list of papers by topic or by authors. By selecting a topic, the list of all the papers in that area is shown in the *paper list* (Figure 1e). It also provides a way to search for specific papers/authors.

Selected authors are shown in the *selected authors* area (Figure 1b). Once authors are added, their papers are shown in the *paper list* and highlighted in the *popularity of topic* view, matched to the author by color coding, which enables users to see in which topic area a particular author fits. For example, Stuart Card has published mainly in the InfoVis area, as seen by the red color coding in the *popularity of topic* view (Figure 1a). We used black when selected authors wrote a paper together.

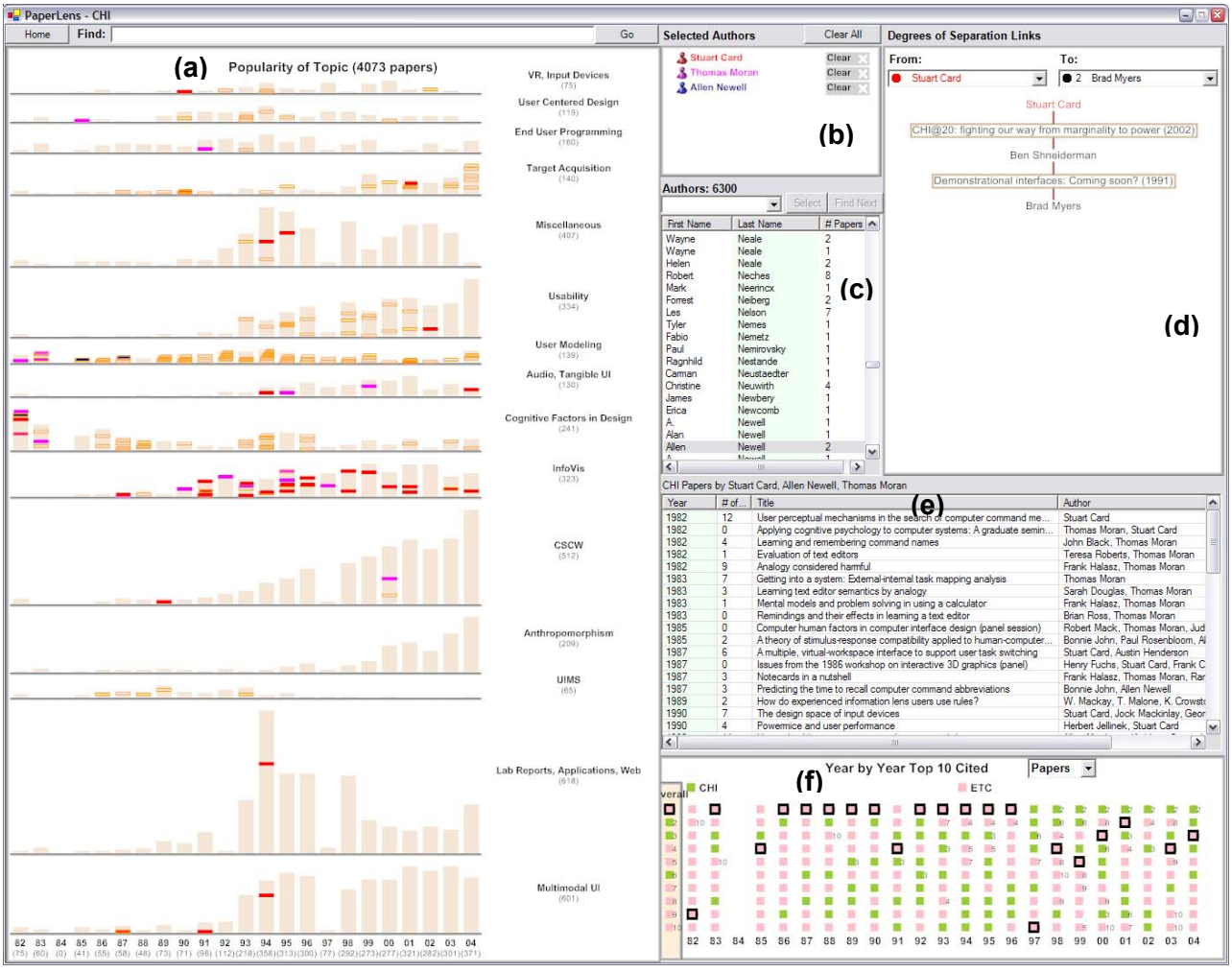

Figure 1. PaperLens tightly couples views across papers, authors, and references and consists of 6 main parts: (a) Popularity of Topic (b) Selected Authors (c) Author List (d) Degrees of Separation Links (e) Paper List (f) Year by Year Top 10 Cited Papers/Authors.

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 see who has published the most. For example, the most prolific author is Brad Myers who has published 41 papers.

One interesting question is "Which papers/authors are most often referenced?" In addition to counting references, we computed them by year and by topic to show trends. When the user selects a topic from the *popularity of topic* view, the *year by year top 10 citations* area (Figure 1f) is filtered to show the frequent citations for that topic area. In this way, the user can quickly discover the influential papers in a particular topic area.

Ranking the frequent citations by author shows frequently cited authors. For End User Programming, Brad Myers was the most frequently cited author. Selecting an author from the *year by year top 10 cited authors* view shows papers that the selected author has published in CHI, and papers that have referenced them, using orange highlighting in the *popularity of topic* area. The user can immediately discover areas most influenced by the selected author.

A co-author collaboration graph is often used to find the relationship between authors [8]. The graph among CHI authors, however, is too fragmented to give useful insights. Instead, we display the shortest path between two authors by co-authorship in the *degrees of separation links* view (Figure 1d). For example, Card and Myers are connected indirectly to each other because they have each co-authored a paper with Shneiderman.

PaperLens was implemented in C# and runs on any standard Windows PC. All the graphical views are implemented with Piccolo.NET, a shared source toolkit that supports scalable structured 2D graphics [2,7].

USER STUDY

A user study was carried out using the InfoVis dataset and the first iteration prototype [4]. Eight researchers (including 1 pilot subject) were recruited. 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 ranged from 24 to 42. The pilot data is included only in the discussion of the usability issues observed.

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

Next, the participants were asked to carry out 16 tasks, which were timed and scored for correctness. 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. Once all of the tasks were completed, users were asked to fill out a questionnaire. All sessions lasted no more than one hour. Participants received a free lunch 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

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. Incorrect answer times were not included in the task time analysis.

Average task times were fast, with only the last task taking much longer (65 seconds, on average). This task was to figure out how many papers Mackinlay and Card published together at InfoVis, which required users to remember that black color coding was used to signify multiple co-authors. Most tasks were performed in less than 20 seconds.

Several usability issues needed to be addressed through design iteration. These issues were prioritized based on how many participants encountered them and the severity

of the issue based on how long the issue delayed finding an answer. The highest priority issues centered on searching for authors: in this prototype a string-based search did not allow the user to search for first or last names separately, and found substring matches anywhere in the name. We addressed this by providing columns for the first and last name, in addition to fixing the way substring matches worked.

Several high priority issues were observed where our system did not behave "symmetrically". If you could launch a paper from one list view, you should be able to open it from any list view. All symmetry issues have been addressed in the redesigned system.

Finally, some users thought the originally separated *degrees of separation list and links* views were "recreational" and took up too much screen real estate. To help alleviate usability issues in this area we combined the list and links views into one view (now called *degrees of separation links*) and allowed the user to pick the degrees of separation between any selected author and related people.

LESSONS LEARNED

We learned two core things in the design and use of PaperLens. The first is that sometimes simple is good. Our initial thoughts were to build a graph visualization tool to show all the data and relationships at once. But we suspected viewing too much information could be overwhelming. Furthermore, there is no efficient way to show topic trends with graph visualizations.

We instead opted for a simpler design with an abstract overview of the full dataset but not with all relationships visible. We also designed around several simple tightly coupled views which provide powerful capabilities together. While these design ideas have appeared before, PaperLens brings them together in a unique fashion. To summarize, key elements of the PaperLens design are: a) an abstract overview; b) multiple small and simple components to best show the different aspects of the data; c) relationships shown through interactivity and tightly coupled components; and d) all visual elements are laid out along axes with well defined metrics.

The second thing we learned pertains to 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. A fisheye technique helped users reveal individual paper titles for a selected year by topic. However, when we tried a similar approach for the CHI data, the height of the rectangle was too small. So, we rendered each rectangle 4 pixels high, and raised highlighted rectangles to the front. However, when several papers are highlighted, rectangles sometimes overlap, causing them to shift one pixel to the right. Overlapping made it difficult to select a paper, so we provided a pop-up list menu showing the papers close to the cursor. The fisheye technique did not work because the number of papers for which we could show titles in one column was less than 70, and we needed to show as many as 150.

Some users were concerned with the number of different colors in the original user interface. 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. We used a single color if the number of selected authors is larger than 4.

FUTURE WORK

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.

ACKNOWLEDGMENTS

We would like to thank the InfoVis 2004 contest chairs, Jean-Daniel Fekete, Georges Grinstein and Catherine Plaisant and the ACM for providing the data. Chris Meek and Susan Dumais helped with clustering solutions. We also thank Aaron Clamage who ported Piccolo from Java to .NET, and the user study participants.

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. InfoVis 2004 Contest: The History of InfoVis http://www.cs.umd.edu/hcil/iv04contest.
4. Lee, B., Czerwinski, M., Robertson, G., and Bederson, B.B. (2004) Understanding Eight Years of InfoVis Conferences Using PaperLens. *Posters Compendium of InfoVis 2004*, 53-54.
5. McLachlan, G.J. and Basford, K.E. (1988) Mixture Models. Marcel Dekker, New York.
6. Nowell, L.T., France, R.K., and Hix, D. (1997) Exploring Search Results with Envision. *Extended Abstracts of CHI 1997*, ACM Press, 14-15.
7. Piccolo.NET. http://www.cs.umd.edu/hcil/piccolo
8. 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.
9. 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. *Posters Compendium of InfoVis 2004*, 51-52.