ThemeExplorer: A Tool for Understanding the History of the Field of Information Visualization

Urs Bischoff, Nicholas Diakopoulos, Felix Loesch, Ying Zhou
Georgia Institute of Technology
{urs, nad, felix, yingzhou}@cc.gatech.edu

ABSTRACT

ThemeExplorer is a tool for exploring a large collection of research papers. Visualization of data uncertainty is one of its key innovations. Several aspects of the dataset are visualized in our focus+context framework. We make use of color and shape to highlight focused documents while constantly visualizing the whole dataset as a contextual overview. Similarity and categories of papers as well as the evolution of categories are visualized in an overview. We employ dynamic-query sliders to support interactive queries, providing users full control of what they want to see. Detailed information about selected papers is shown in the detail pane.

THE VISUALIZATION

ThemeExplorer uses several well known visualizations and techniques combined with our own additions, all built within an overview and detail framework, allowing the user to explore the same data from different viewpoints (Figure 1). The landscape overview provides a static view of the document corpus within the past ten years of the InfoVis conference by using a "thematic terrain" metaphor that displays the primary themes and their relative prevalence in one single view, as described by Wise et al. [1]. The user can instantly see which theme has the largest number of documents by looking at the height of the mountains. Each document is represented by a dot on the landscape. One of the key innovations of our system is the visualization of data uncertainty. Referenced documents that could not be mapped on the landscape due to the lack of an abstract are displayed in the black uncertainty region above the landscape (Figure 2). The uncertainty region extends the contextual overview because all references are shown and enables the user to explore these documents which would not have been possible without the uncertainty region.

Another powerful feature of our system is the superimposed tree visualization on the landscape (Figure 2); we use it to show references and citations of documents. The tree visualization is not only useful to find quick information about referenced articles, but it is also useful for browsing the data. We found it very valuable for tasks like "finding the article that is cited most". Differences in the number of references can be easily

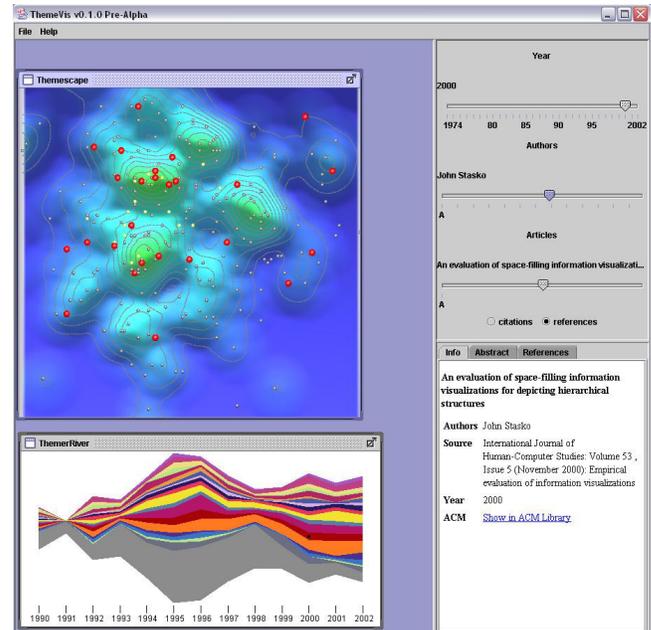perceived. The coloring of the lines supports a quick recognition of the source article.

Figure 1: ThemeExplorer is a hybrid of several known visualizations and techniques built within an overview and detail framework.

Although the landscape visualization is very powerful, it cannot show certain important aspects of the data set such as the evolution of research areas over time. We use the ThemeRiver technique [2] (see left bottom sub window in Figure 1) to show the evolution of research areas over time. The research areas in the document collection are represented by a "river" that flows left to right through time. The river widens or narrows to depict changes in the collective strength of selected research areas in the underlying documents. The total number of documents for a year can be seen by looking at the overall width of the river. ThemeExplorer helps to visualize uncertainty in the dataset by adding two grayish sub-rivers (on the bottom side of the river). Thus, documents that are not similar to any theme and documents that do not have an abstract are shown in these sub-rivers. The different topics of the area of information visualization are based on key words from

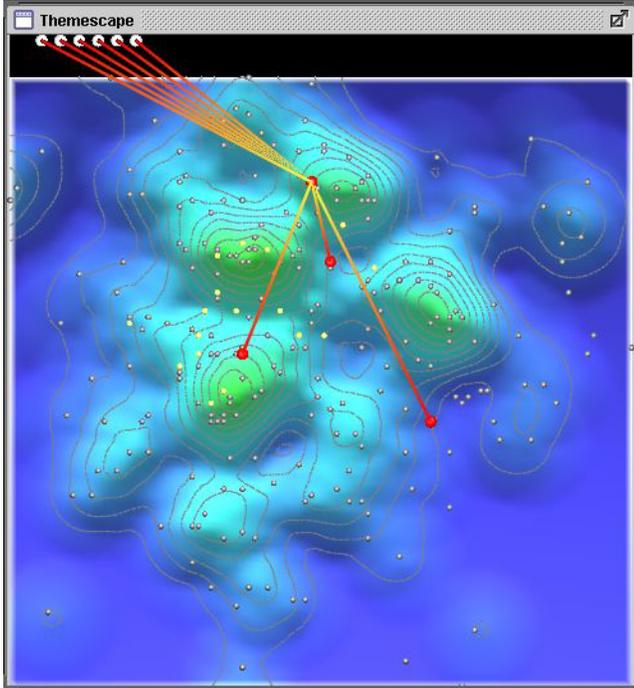an InfoVis course taught by Prof. John Stasko at Georgia Tech.

Figure 2: Static landscape overview with superimposed tree visualization. The selected article is shown in the center and referenced articles are connected by lines. Articles that are not mapped on to the landscape are visualized as white dots in the black uncertainty region. By switching to the citations mode the referencing articles highlighted and connected instead of the referencing articles.

USER INTERACTION

The key elements of the user interaction are dynamic queries, brushing, focus+context, zooming and panning. The brushing technique allows the user to select or highlight an element in one view, generating highlights of the element in another view. For example, in ThemeExplorer, the user may select a particular theme from ThemeRiver by clicking on a colored river; all documents under this theme will be highlighted on the landscape. Each visualization in our framework itself is already powerful, but the combination of the different views on the same data and brushing allows the user to perform complex tasks that are not possible with one single visualization.

Another important strength of ThemeExplorer is the dynamic query capability. The three provided dynamic query sliders (see right part of Figure 1) enable users to filter what they want to see by year, author and articles. The dynamic query sliders directly affect the landscape and the ThemeRiver as well as the detail information window on the right side. Once a document is selected, more detailed information is presented below the sliders, including abstract, references, citations and a hyperlink to

the full paper. The dynamic query sliders are logically connected and therefore affect each other. For example, users may first select by year to narrow down the search. "Authors" and "Articles" are automatically updated to reflect the year selected. Also all articles that satisfy the dynamic query are highlighted on the landscape. While brushed data is highlighted and displayed, the non-brushed data is dimmed or grayed out in the background, and serves as context for the brushed data. A number above the sliders is constantly updated and indicates how many articles are currently "covered" by the sliders. We determined the way that a slider mutually affects other sliders by presenting different design ideas to a user group. Dynamic query operations also affect the ThemeRiver insofar as small dots will be placed in the sub-rivers corresponding to the themes that are similar to a selected document.

We support zooming and panning which allows the user to have a closer look at a certain area of the landscape.

IMPLEMENTATION
ThemeExplorer is fully implemented in Java. Automated text analysis is used as a pre-processing step for the visualizations. A term feature vector based on the abstract text is produced for each document with an abstract. Based on a standard cosine distance matrix, the articles are then projected on a plane and categorized into different research areas.

LIMITATIONS
One major limitation of our system is that the overview visualization (height field of the landscape and the ThemeRiver) is static. This means that the landscape is always based on the similarity of the textual content and the ThemeRiver depicts the evolution of topics over time. This makes ThemeExplorer a little bit less powerful for certain tasks like analyzing the relationships between authors. For this specific task one has to use the alpha slider that selects a certain author. Then, detail information for articles of that author is displayed; for example, the names of co-authors are shown. Furthermore, referenced articles are visualized on the landscape which can also be helpful to analyze author relationships.

REFERENCES

1. J. A. Wise, J. J. Thomas, K. Pennock, D. Lantrip, M. Pottier, A. Schur, V. Crow, Visualizing the non-visual: spatial analysis and interaction with information from text documents, Proceedings of the 1995 IEEE Symposium on Information Visualization, p.51, October 30-31, 1995, Atlanta, Georgia

2. Susan Havre, Beth Hetzler, Lucy Nowell, ThemeRiver: Visualizing Theme Changes over Time, Proceedings of the IEEE Symposium on Information Vizualization 2000, p.115, October 09-10, 2000