ThemeRiver: Visualizing Theme Changes over Time

Susan Havre, Beth Hetzler, and Lucy Nowell
Battelle Pacific Northwest Division
Richland, Washington 99352 USA
1+509+375-6948
{susan.havre | beth.hetzler | lucy.nowell}@pnl.gov

Abstract

ThemeRiver™ is a prototype system that visualizes thematic variations over time within a large collection of documents. The "river" flows from left to right through time, changing width to depict changes in thematic strength of temporally associated documents. Colored "currents" flowing within the river narrow or widen to indicate decreases or increases in the strength of an individual topic or a group of topics in the associated documents. The river is shown within the context of a timeline and a corresponding textual presentation of external events.

Keywords: visualization metaphors, trend analysis, timeline

1. Introduction

In exploratory information visualization, one goal is to present information so that users can easily discern patterns. Patterns reveal trends, relationships, anomalies, and structure in the data, and may help users

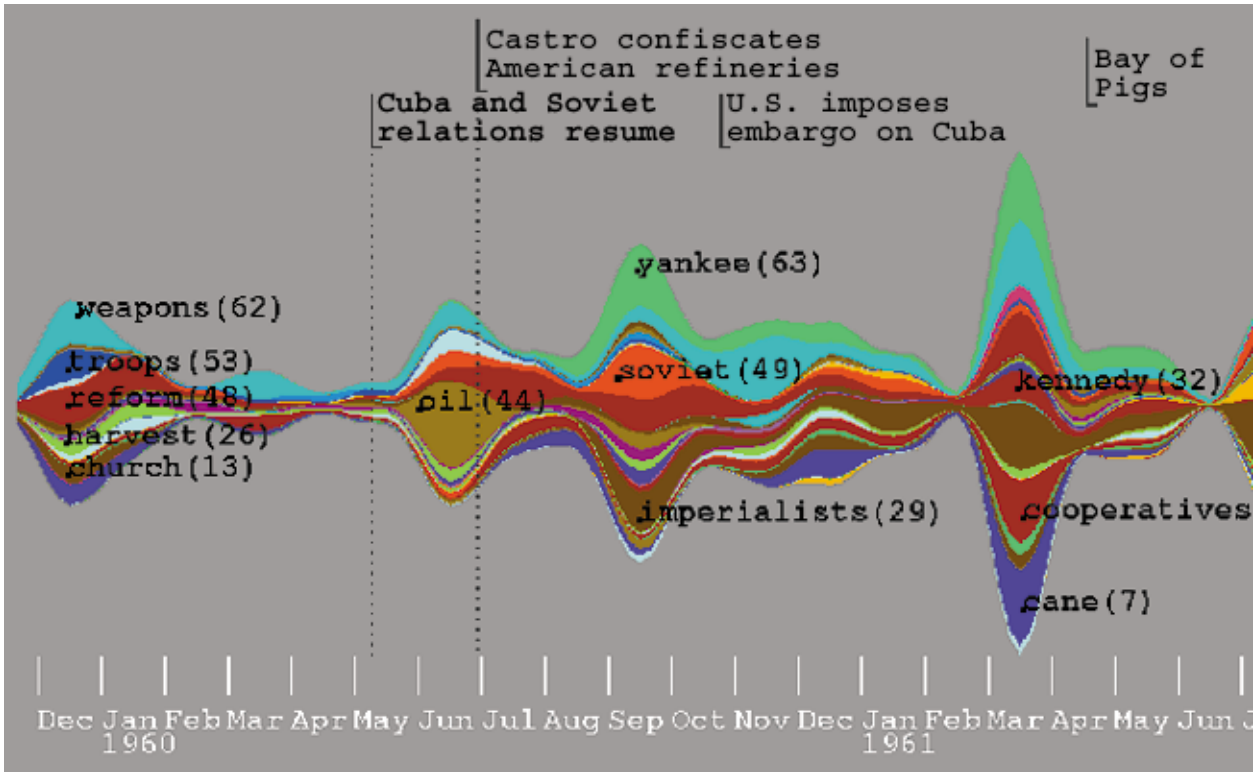

Figure 1: ThemeRiver™ uses a river metaphor to represent theme changes over time.

confirm knowledge or hypotheses. Perhaps more importantly, they also raise unexpected questions leading users to new insights. The challenge is to create visualizations that enable users to find patterns quickly and easily. ThemeRiver, shown in Figure 1, is a prototype system designed to reveal temporal patterns in text collections.

Information visualization systems such as Envision [13], BEAD [1], LyberWorld [3, 4] and SPIRE [18] represent each document or group of documents with a glyph or icon, portraying various document attributes. Various methods have been explored for showing change over time in document-centric visualizations. See Section 3 below.

However, a user may be less interested in documents themselves than in theme changes within the whole collection over time. For example, how did Shakespeare's themes change during various periods of his life or in relation to contemporary events? Such information is difficult, if not impossible, to glean from most visualizations. A visualization that focuses on themes, rather than documents, could be more useful for such exploration.

ThemeRiver provides users with a macro-view of thematic changes in a corpus of documents over a serial dimension. It is designed to facilitate the identification of trends, patterns, and unexpected occurrence or non-occurrence of themes or topics. In our prototype, we use time as the serial dimension. We provide contextual information through a timeline and markers for co-occurring events of interest. Figure 1 shows a sample ThemeRiver visualization. This paper describes the design of ThemeRiver, walks through a sample information exploration session, and discusses results of formative usability testing.

2. Design

Our major design goal was to provide a visualization of theme change over time. Consider using a histogram to visualize these changes. In a histogram (such as the one shown in Figure 2), each bar represents a time slice, and color variations and size within the bar represent the relative strength of themes specific to that slice. However, understanding the histogram requires users to work at integrating the themes across time because the bars are anchored to a baseline and the position of a particular theme within the bars may vary considerably.

Like a histogram, ThemeRiver uses variations in width to represent variations in strength or degree of representation. However, it connects the strength values in adjacent time slices with smooth and continuous curves. The horizontal flow of the river represents the flow of time. Colored currents that run horizontally within the river represent themes. Each vertical section of the river corresponds to an ordered time slice.

The width of each current changes to reflect the thematic strength for each time slice. For example, in Figure 1 the theme "soviet" increases in relative strength in June 1960 as indicated by the widening of the upper bright orange current. "Soviet" loses relative strength in July and August; thus the same current narrows in the next two time slices. "Soviet" then increases significantly in relative strength in September; the current widens proportionately.

Currents maintain their integrity as a single entity over time. If a theme ceases to occur in the documents for a period of time and then recurs, the current likewise disappears and then reappears. Consistent color and relative position to other themes make theme currents easy to recognize. In Figure 1, the lower purple band depicts the changes in relative strength of the theme "cane." The "cane" current occurs grows and shrinks over time; "cane" occurs most strongly in March 1961.

We believe that ThemeRiver's continuous curves have much to do with its usability. The Gestalt School of Psychology [8], founded in 1919 in Germany, theorized that with perception, "the whole is greater than the sum of the parts." Simply put, during the perception process humans do not organize individual, low-level, sensed elements, but sense more complete "packages" that represent objects or patterns. In his recent book [6], Hoffman presents a compelling discussion of how our perceptual processes identify curves and silhouettes, recognize parts, and group them together into objects. Numerous aspects of the image influence our ability to perceive these parts and objects, including similarity, continuity, symmetry, proximity, and closure. For example, it is easier to perceive objects that are bounded by continuous curves than those that contain abrupt changes [17].

The vertical proximity of the river currents makes it easy for users to judge the relative width of currents and thus the relative strength of the themes. Similarly, symmetry around the horizontal axis of the river, a current, or group of currents makes it easier for users to perceive flow patterns and changes. Widths of currents combine to show cumulative widening and narrowing, representing changing strength for the selected set of themes as a whole.

Values for theme strength can be calculated various ways. For example, they might represent the number of documents containing the word. Because the river loses its continuity and structure if there are too few or too many themes, we created several theme subsets for exploration.

We have implemented a proof-of-principle prototype and used it to explore data from multiple sources. Figure 1 portrays data from a collection of speeches, interviews, articles, and other text associated with Fidel Castro. The visualization includes the river, a timeline below the river, and markers for related historical events along the top. With ThemeRiver, users may

- display topic and event labels
- display time and event grid lines
- display the raw data points
- choose among drawing algorithms for the currents and river.

Users may also display the associated time or theme name by simply moving the mouse across the image. In addition, users may pan and zoom to see other time periods or parts of the river and to see more detail or broader context. In this sample data set, we found several interesting correspondences between themes and events, such as the expansion of the "oil" theme just before Castro confiscated American oil refineries (see Figure 1).

3. Related Work

Many systems include features for viewing time. One common method is to show discrete time slices. For example, in the Spatial Paradigm for Information Retrieval and Exploration (SPIRE) Galaxy visualization [18], users may choose to progressively step through time, showing only the icons for documents originating within each specified time period. Another common approach is to show time as an attribute of documents, as done in the Virginia Tech's Envision system, which lets users map various metadata values, including date, to x-axis, y-axis, or color, shape, or size graphical encodings [13].

More similar to ThemeRiver in intent are systems that focus directly on time. The LifeLines system, developed jointly by the University of Maryland and IBM, has been used to visualize medical records and juvenile criminal records [14, 15]. The visualization displays time along the x-axis and uses the y-axis to categorize events. Bars depict duration for a given event, and graphical attributes such as color show event attributes. TmViewer uses a similar approach, adding the ability to show parent-child relationships with lines between related time bars [10]. The DIVA system [12] uses animation to show how particular measured values change in relation to the temporal flow of a video. To help groups collaborating to create a document or other artifact, the Timewarp system developed at Xerox PARC [2] lets users view and edit multiple timelines of the changing state of that artifact. The metaphor used is similar to a state diagram, with lines connecting state nodes and branches. Additional work on timelines includes Karam's [7] and Kullberg's [9].

We know of no other systems that use the river metaphor to depict the passage of time. However, Tufte [16] presents a similar idea in an artist's illustration showing trends in music. In that illustration, width represents sales and proximity indicates influence of preceding styles. Our work differs in several aspects, such as the use of color, the inclusion of contextual events, and the ability to generate the visualization automatically from a potentially very large collection of documents.

4. Usability Evaluation

Early in ThemeRiver's development, we carried out a simple formative usability evaluation with two users. Questions we wanted to answer with this evaluation included

- Do users understand the metaphor?
- Can they identify themes that are more often discussed?
- Does the visualization help them raise new questions about the data?
- Do they interpret details of the visualization in ways we had not expected?
- How does their interpretation of the visualization differ from that of a histogram showing the same data?

The data were the Castro collection described above, focusing on the years 1960-1963. We represented the same data both in ThemeRiver and in a histogram that we created using a spreadsheet. (See Figure 2.) We made the content of the histogram as similar as possible to ThemeRiver's. For example, the histogram depicted thematic content by months, using the same values that drive ThemeRiver. The month timeline was shown along the bottom and we added an event line to the histogram like the one in ThemeRiver.

Usability evaluation began with a brief explanation of the purpose of the session, followed by an introduction to the data. Both participants viewed the data in both visualizations; one participant started first with the

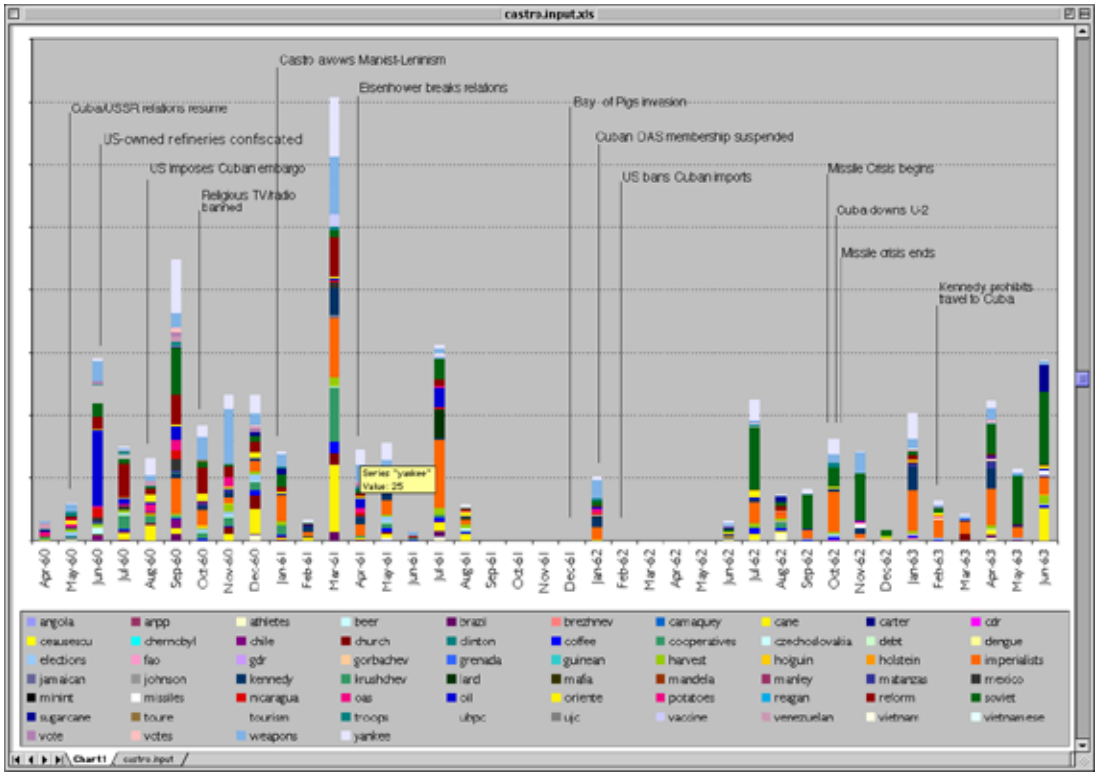

Figure 2: Like ThemeRiver™ in Figure 1, this histogram uses the Castro collection data and depicts changes in thematic content over time.

histogram and one with ThemeRiver. We asked each participant questions about what they observed in each display.

Examples of specific questions include

- In July '62, what are the three most discussed themes?
- Where is a new theme introduced?

Examples of more general questions include

- What looks interesting here – what do you want to explore?
- How would you like to change or manipulate the view?

We captured verbal protocol during this discussion. At the end, we asked participants to complete a short questionnaire, with feedback about the visualization and possible enhancements.

From the verbal protocol and from user behavior, we observed that the users had no difficulty in understanding the metaphor. They were able to identify themes that were strongly represented and able to understand the relationship between the width of the current and theme strength. The visualization also triggered questions about the reasons behind certain theme strengths and patterns. For exploratory visualizations, this is a good result; we believe that a visualization should help the user identify questions of interest to explore.

Questionnaire responses showed that users found ThemeRiver easy to understand. They also found ThemeRiver useful, particularly for identifying macro trends. They told us that it was less useful for identifying minor trends because the curves tend to de-emphasize very small values. We asked about the value of the river metaphor, and users rated it highly as well. They observed that the connectedness of the river helped them follow a trend more easily over time than in the histogram; this result is compatible with the perception principles described by Ware [17].

Users liked some features of the histogram and recommended adding them to ThemeRiver. One such feature is the ability to see numeric values that drive the histogram and river currents. One user expressed more trust in the histogram, because she "knew" that the bars were exactly the data values, whereas she was not sure exactly what the data values were in ThemeRiver. Her point is a valid one, especially because the curved lines

of ThemeRiver do require that we interpolate between data points to produce the curves. We have added the capability for users to see the exact data points on demand.

Although users liked the abstraction to the whole collection and thus away from individual documents, both users suggested adding features to access documents if desired. They wanted the ability to see the total number of documents during any time period and to get the text of each document on demand. They wanted to select a current and see the documents that contributed to it.

Users also wanted the ability to reorder the theme currents. Options they discussed included user-defined ordering and ordering by correlation, so that themes appearing together in the documents would be nearby in the river.

5. Interactions and Sample Usage

Based on usability evaluation results, we added a number of features to combine the best of both the river metaphor and histogram capabilities. This section presents a sample usage scenario, illustrating the capabilities of the current version.

We used ThemeRiver to explore the 1990 Associated Press (AP) newswire data from the TREC5 distribution disks, a set of over 100,000 documents (see Figure 3). To explore the selected themes in this collection, a user might begin with a high-level survey of the visualization by panning along the course of the river. The user might look for wider currents that signal heavy use of a topic, such as the one for "baghdad" in Figure 3. Changes in the color distribution of the river signal changes in themes. We see such a change in August 1990, when the "kuwait" current, which had vanished in late July, suddenly appears and rapidly widens. The user could also look for narrow currents in the river that signal relatively light use of particular themes.

In an earlier paper, Hetzler et al. [5] explored the AP data set with a variety of our visual analysis tools, focusing on large theme changes surrounding the Iraqi invasion of Kuwait on August 2. ThemeRiver also reflects these large theme changes. Near the right side of Figure 3, we see several currents that expand dramatic-

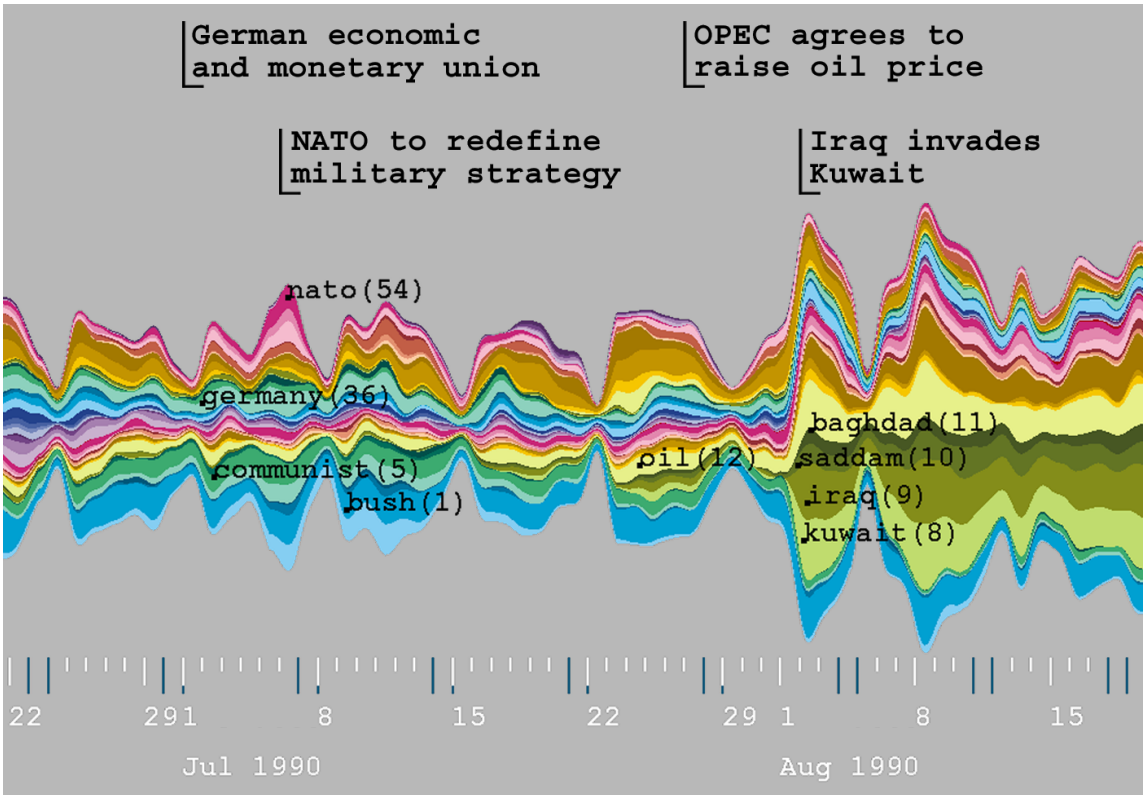

Figure 3: AP data from July - August 1990. A wide current in the river indicates heavy use of a topic, while changes in color distribution correlate to changes in themes.

ally at the time of the invasion, which is shown on the event line above the river. Labels have been turned on for currents representing the themes "kuwait," "iraq," "saddam," and "baghdad." ThemeRiver reveals some additional detail not noted in the earlier study. The theme "oil," which is persistent across the image, also expands noticeably at this time. The themes of "kuwait," "iraq," and "saddam" show up in small bursts before the invasion but are not persistent. News stories corresponding with these bursts covered the verbal conflicts leading up to the invasion. This distinction between persistent and bursty themes is one advantage that ThemeRiver provides over document-centric visualizations.

During late June and throughout July 1990, the themes appear relatively consistent. A user interested in the more prominent themes might turn on theme labels as shown in Figure 3 to discover that the main themes represent "bush" (President Bush), "germany" (the reunification discussions), and "communist." Some smaller variations in theme are also apparent, such as the widening of the "nato" band, related to the NATO decision to redefine their military strategy.

Figure 4 shows the ThemeRiver from earlier in the summer of 1990. In late May, a large change in theme strength is shown, this time not matching any previously identified events. Some of the larger currents here are "gorbachev," "bush," and "summit." This might suggest that Bush and Gorbachev both attended a summit. Viewing the pertinent news documents from that time, we found that a four-day summit meeting took place in Washington among several world leaders, including Bush and Gorbachev.

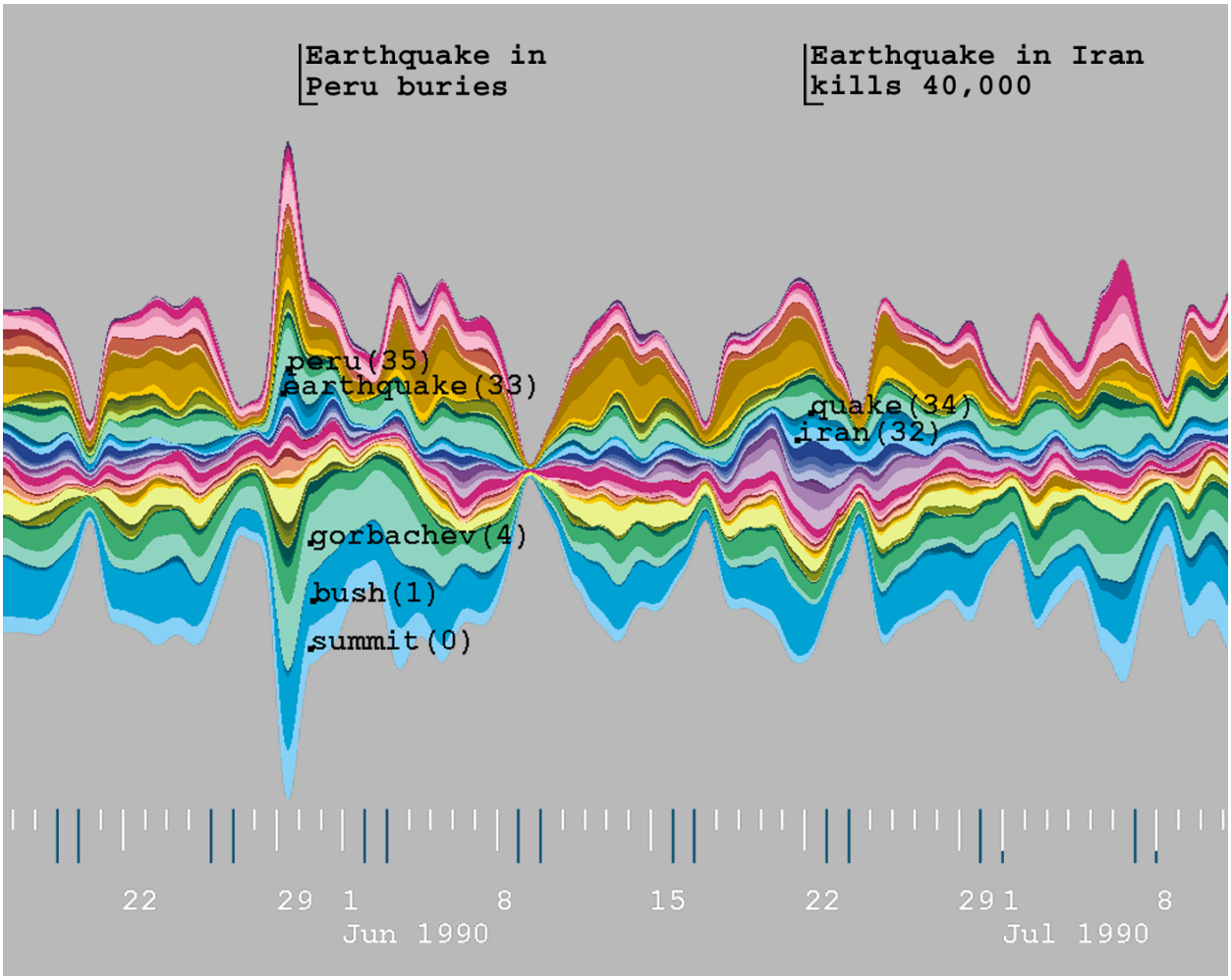

Figure 4: ThemeRiver™ of AP data from June - July 1990 identifies very different events from those revealed immediately afterwards (Figure 3).

Some more subtle changes can also be seen in Figure 4. For example, a small current near the middle of the river expands slightly near the beginning of June and again near the end of the month. This is the current for "earthquake." The wider areas correspond with the quakes in Peru and Iran respectively.

In each of the figures shown so far, there are portions of the river that are extremely narrow overall. In fact, for the AP rivers (Figures 3 and 4), the river seems to narrow quite frequently. On closer inspection, we see that the narrow spots correspond with Sundays. Because the river contains only a subset of the themes in the collection, we do not know at this point whether the news is generally lighter on Sunday or whether other topics dominate on that day. This uncertainty is one of the points that came up early in user testing. In response, we have added a feature allowing the user to show a histogram representing the total number of documents in a given time slot, along with the portion represented by the themes in the river (see Figure 5). With this histogram, it is apparent that in general fewer news stories are released on Sunday than on other days of the week.

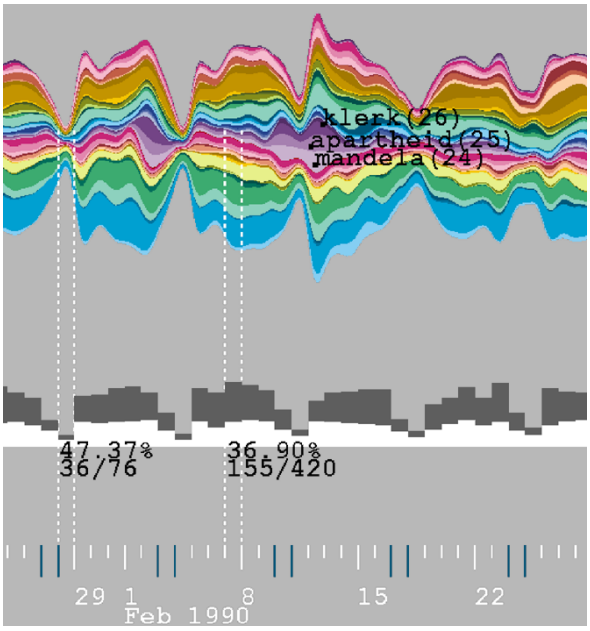

Figure 5: The addition of a histogram to ThemeRiver™ reveals that news is light on Sundays, not that themes shift.

Sometimes users may want to compare theme changes in one set of documents to those in another set; alternatively, they may wish to partition a collection based on metadata and compare the themes in the two partitions as separate rivers. Figure 6 shows two parallel rivers: the lower river shows AP news stories from Washington, D. C. and the upper river shows the news stories from New York. Some differences in major themes are immediately apparent. The Washington themes emphasize Bush, the Senate, and the Supreme Court. The New York stories show a major growth in the themes "apartheid" and "mandela"; this corresponds with the visit of Nelson Mandela to the US. He arrived first in New York, where he spent several days before proceeding to Washington.

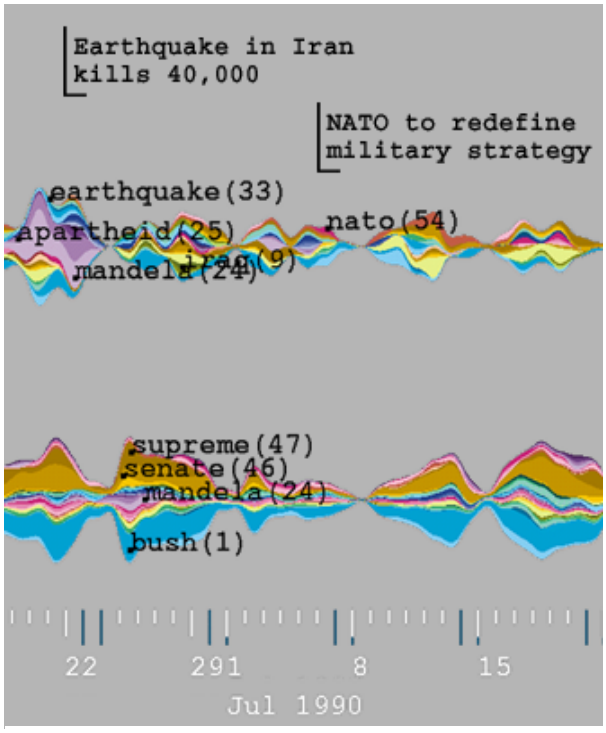

Figure 6: Parallel rivers let users compare AP data from Washington, D.C. and New York from the same time period.

6. Discussion and Design Challenges

Ideally, a visual metaphor facilitates discovery by presenting data in an intuitive, easy way that is consistent with the user's perceptual and cognitive abilities. Lakoff and Johnson [11] argue that metaphors are wired into our understanding of particular concepts, using evidence from common linguistic expressions. One example they cite is the many English expressions that imply that Anglo-Americans understand time in terms

of motion relative to ourselves. Some expressions characterize time as moving (e.g., "the time will come," "don't let the opportunity pass"), while others imply that people are the ones moving through time (e.g., "as we go through the years"). From formative usability evaluation and anecdotal feedback, we have observed that the river metaphor is intuitive and easy to understand. We believe the river metaphor of theme currents changing over time gets part of its strength from this cultural understanding.

Focusing on themes rather than documents changes issues of scalability. ThemeRiver visualizations have little dependence on the number of documents represented. For example, if theme strength is determined by the number of documents containing each theme word, a single pass through the collection is needed to calculate the values, which may be displayed similarly regardless of collection size. On the other hand, the number of currents that can be reasonably included in a single river is limited. Options for addressing this issue include grouping through color families, as suggested in Figure 7, or using each current to represent a set of themes rather than a single theme.

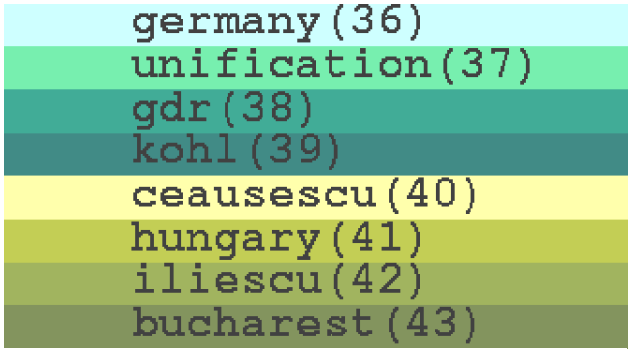

Figure 7: Tracking related themes is simplified by assigning them to the same color family. This ensures related themes appear together and are identifiable as a group.

Color choices pose an interesting design challenge. Color perception depends on local contrast. However, because themes come and go, it is impossible to predict which colors will be adjacent at any given time. Moreover, we want to show a relatively large number of themes in the river and still achieve acceptable discriminability. Currently we are exploring a solution suggested during formative usability evaluation: sorting themes into related groups and displaying each group with a color family. Figure 7 shows a portion of our color legend with such an ordering, which emphasizes changes in related themes and may make it easier to understand relationships among them.

A key cognitive advantage of the river metaphor over a simple histogram lies in the curving continuous lines that define the boundaries between topic currents. But it is also important that the visualization not mislead users. Because dates are not continuous data, we must approximate the true boundaries by interpolating between discrete data points. As long as the resolution of the data is sufficient, ThemeRiver provides an overview that meets our criteria for intuitiveness, ease of use, and integrity. If the user zooms in farther than the data resolution supports, the "truthfulness" approximated by the interpolated lines is questionable.

While the resolution of data forces a lower limit on the level of zoom, we can deal with the problem of "too much" resolution by combining time slices. That is, as the user zooms out, we can increase the amount of time per time slice and combine theme weights. In this way, we can maintain a suitable level of truthfulness without slowing the rendering speed to a crawl by trying to draw more detail than necessary.

With interactive visualizations, calculation and drawing speeds are important. For the current features of ThemeRiver, it is sufficient to calculate the drawing points on startup and then recalculate only after a configuration change. Nevertheless, a fast, efficient algorithm is needed. We are investigating curved-line algorithms and ways to speed up both the calculations and the rendering.

7. Conclusions

ThemeRiver is a demonstration prototype, developed to test the value of the metaphor. We are continuing to add interaction capabilities to it. We also need to develop ways to build the event timeline automatically and to provide more flexibility in selecting and ordering the theme currents. From formative usability evaluation, we learned that users want to know more about the context of the river and want to access the documents that contribute to it at a particular point in time.

We conclude that ThemeRiver is potentially valuable for information analysts and plan to develop it into a full production system.

8. Acknowledgments

We gratefully acknowledge the contributions of our colleagues at Battelle to the development and testing of the ThemeRiver visualization. Special thanks for contributions to this paper go to Grant Nakamura, Alan Willse, Sharon Eaton, Wanda Mar, and Dan Donohoo. Battelle Memorial Institute's Information Synthesis Platform funded this research.

9. References

1. D. Brodbeck, M. Chalmers, A. Lunzer, and P. Cotture, "Domesticating Bead: Adaptiing an Information Visualization System to a Financial Institution," *Proceedings of InfoViz '97*. IEEE Computer Society, Los Alamitos, CA, 1997, pp. 73-80.
2. K.W. Edwards and E.D. Mynatt, "Timewarp: Techniques for Autonomous Collaboration," *Proceedings of CHI'97*, Association for Computing Machinery, Inc., 1997, pp. 218-225.
3. M. Hemmje, "LyberWorld: a 3D Graphical User Interface for Fulltext Retrieval," *Conference Companion on Human Factors in Computing Systems*, 1995, pp. 417 - 418.
4. M. Hemmje, C. Kunkel, and A. Willett. "LyberWorld - a Visualization User Interface Supporting Fulltext Retrieval," *Proceedings of the 17th Annual International ACM-SIGR Conference on Research and Development in Information Retrieval*, 1994, pp. 249 -259.
5. B. Hetzler, P. Whitney, L. Martucci, L., and J. Thomas, "Multi-faceted Insight Through Interoperable Visual Information Analysis Paradigms," *Proceedings of IEEE Symposium on Information Visualization, InfoVis '98*, 1998, pp.137-144.
6. D.D. Hoffman, Visual Intelligence: How We Create What We See, W.W. Norton & Company, Inc., New York, 1998.
7. G.M. Karam, "Visualization Using Timelines," Proceedings of the 1994 International Symposium on Software Testing and Analysis, 1994, pp. 125-137.
8. K. Koffka, (1935), *Principles of Gestalt Psychology*, Harcourt-Brace, New York, 1935.
9. R.L. Kullberg, "Dynamic Timelines: Visualizing the History of Photography," *Proceedings of CHI '96*, 1996, pp. 386-397.
10. V. Kumar and R. Furuta, "Visualization of Relationships," *Proceedings of Hypertext 99,* ACM Press, Darmstadt, Germany, 1999.
11. G. Lakoff and M. Johnson, *Metaphors We Live By*. University of Chicago Press, Chicago, 1983.
12. W. Mackay and M. Beaudouin-Lafon, "Diva: Exploratory Data Analysis with Multimedia Streams," *Proceedings of CHI'98*, 1998, pp. 416-423.
13. L.T. Nowell, R.K. France, D. Hix, L.S. Heath, and E.A. Fox, "Visualizing Search Results: Some Alternatives to Query-Document Similarity," *Proceedings of SIGIR '96*, ACM Press, Zurich, 1996, pp. 67-75.
14. C. Plaisant, D. Heller, J. Li, B. Shneiderman, R.J. Mushinlin, and J. Karat, Visualizing Medical Records with LifeLines. *CHI '98 Summary*, 1998, 28-29.
15. C. Plaisant, B. Milash, A. Rose, S. Widoff, and B. Shneiderman, "Lifelines: Visualizing Personal Histories," *Proceedings of CHI '96*, Association for Computing Machinery, Inc, 1996, pp. 221-227.
16. E.R. Tufte, Visual Explanations: Images and Quantities, Evidence and Narrative, Graphics Press, Cheshire, CT, 1997, 90-91.
17. C. Ware, Information Visualization: Perception for Design, Academic Press, San Diego, 2000.
18. J.A. Wise, J.J. Thomas, K. Pennock, D. Lantrip, M. Pottier, A. Schur,, and V. Crow, "Visualizing the Non-Visual: Spatial Analysis and Interaction with Information from Text Documents," S.K Card, J.D. Mackinlay, and B. Shneiderman, (editors.), *Readings in Information Visualization: Using Vision to Think*, Morgan Kaufmann, San Francisco, 1999, pp. 442-45