ThemeRiver: Visualizing Thematic Changes in Large Document Collections

Susan Havre, *Member*, *IEEE Computer Society*, Elizabeth Hetzler, Paul Whitney, and Lucy Nowell

Abstract—The ThemeRiver visualization depicts thematic variations over time within a large collection of documents. The thematic changes are shown in the context of a time line and corresponding external events. The focus on temporal thematic change within a context framework allows a user to discern patterns that suggest relationships or trends. For example, the sudden change of thematic strength following an external event may indicate a causal relationship. Such patterns are not readily accessible in other visualizations of the data. We use a river metaphor to convey several key notions. The document collection's time line, selected thematic content, and thematic strength are indicated by the river's directed flow, composition, and changing width, respectively. The directed flow from left to right is interpreted as movement through time and the horizontal distance between two points on the river defines a time interval. At any point in time, the vertical distance, or width, of the river indicates the collective strength of the selected themes. Colored "currents" flowing within the river represent individual themes. A current's vertical width narrows or broadens to indicate decreases or increases in the strength of the individual theme.

Index Terms—Visualization, metaphor, trend analysis, time line.

1 Introduction

Today's knowledge workers must make sense of huge amounts of unstructured textual data in the form of document collections. Exploring multiple visual presentations, or visualizations, of the data often helps a user make sense of a collection. Each visualization or some combination of visualizations may lead to important insights and/or a better global understanding of the collection.

Some visualizations of document collections are based on features derived from the textual content of the documents or groups of documents. Such visualizations depict relationships among the documents or groups of documents based on the derived features. The relationships are often indicated by the proximity of document icons in a two or three-dimensional projection. Examples of such a view include the SPIRE (Spatial Paradigm for Information Retrieval and Exploration) Galaxies visualization [1] and Bead [2].

However, users may be less interested in the documents themselves than in the thematic content of the document collection as a whole. Some approaches for portraying collection themes include a landscape metaphor, as in SPIRE's ThemeView visualization [1] and self-organizing maps, such as in Lin [3]. These visualizations allow the user to identify a collection's thematic content, the relative strength of themes, and relationships among themes in a collection.

Time order is an important type of relationship among documents that is important for some analytical tasks. For many document collections, time is a critical attribute. For example, consider the importance of the date and time for understanding a collection of news wire stories, for tracking the evolution of technology through patents and publications, and for portraying discussion flows during a meeting. Time is a natural dimension for humans and is easily interpreted in graphic displays [4], such as time lines. Viewing changes in the thematic content of a body of work over time provides a different and powerful visualization for finding trends and relationships.

The ThemeRiver visualization, shown in Fig. 1, depicts thematic variations in a collection of visualization patents from one company over several years. The thematic changes are presented in the context of a time line, shown at the bottom of the figure. This combination allows a user to discern patterns in individual themes and among multiple themes relative to time. These patterns may reveal trends, relationships, anomalies, and structure in the data. In Fig. 1, for example, the variation of the company's visualization patent activity is obvious. Each period of reduced patent applications is followed by a period of increased patent applications. Focusing on the third bulge, "histogram" (the yellow-gold current above the center), "predictive" (the light blue-green current near the bottom), and "inference" (the dark blue-green current at the bottom) appear for the first time as patent themes in 1994. "Paradigm" (the rusty-red current in the top quarter) was a big theme in 1995. "Internet" (lavender current near the top right), "trends" (light lime current near the center right), "gesture" (light blue-green current near the center right), and "dialogue" (light blue current below center) were new themes in 1996 patents.

Such patterns may confirm or refute the user's knowledge or hypotheses about the collection. Perhaps more importantly, patterns that are identified but that cannot be explained raise questions, leading users to new insights. Patterns that are expected but that cannot be found also

• *The authors are with the Battelle Pacific Northwest Division, Richland, WA 99352.*
E-mail: {susan.havre, beth.hetzler, paul.whitney lucy.nowell,}@pnl.gov.

Manuscript received 17 Apr. 2001; accepted 10 July 2001.
For information on obtaining reprints of this article, please send e-mail to: tvcg@computer.org, and reference IEEECS Log Number 114502.

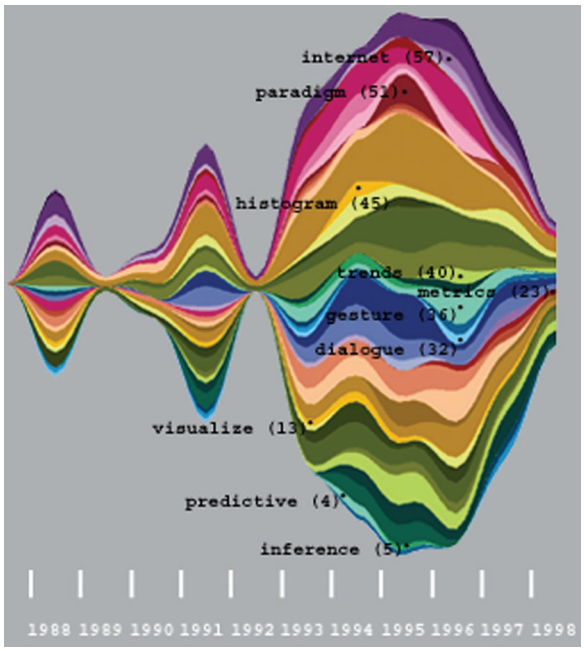

Fig. 1. ThemeRiver depicts thematic changes over time in a collection of patents from one company.

raise potentially important questions. These patterns are not readily accessible in other visualizations of the data.

For another example, consider the complete speeches and interviews of a candidate who is running for an important political office. The composite thematic content can be explored using SPIRE's ThemeView [1]. What are the candidate's major themes? How do these themes relate to one another? What do the choice of themes and their implied relationships suggest about the candidate's views?

ThemeRiver can be used to explore the same data set to look for theme changes over time. Do themes change often? Are they used consistently over time? Which themes occur together? It might be even more interesting to compare speeches and interviews from two opponents. Do the candidates address the same themes? Do the candidates appear to interact or do they ignore each other? Does one candidate always initiate and the other always respond? Such information is difficult, if not impossible, to glean from most visualizations.

In the following sections, we discuss our design goals for ThemeRiver. We briefly describe the ThemeRiver prototype and then walk through a sample information exploration session. We present implementation details of the prototype, such as the data design and curve calculations. Discussion of the results of formative usability testing concludes with a description of changes we made as a result of the testing. Finally, we discuss specific design challenges, conclusions, and plans for future work.

2 Design Goals

In designing new visualizations, one of our goals is always to enable users to find patterns quickly and easily, drawing on the power of the human perceptual system. We use familiar metaphors to help users more easily comprehend the data presentation. We design the graphic layout to leverage a user's ability to quickly assimilate information, drawing on research in cognitive science and psychophysics. The addition of useful contextual information, such as ThemeRiver's time lines and event annotations, allows the user to connect the patterns in content to events or time intervals such as seasons.

2.1 Metaphor

Ideally, a visual metaphor facilitates discovery by presenting data in an intuitive way that is consistent with the user's perceptual and cognitive abilities. Lakoff and Johnson [5] 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 themselves. Some figures of speech characterize time as moving (e.g., "the time will come" and "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"). We believe the river metaphor of theme currents changing over time derives part of its strength from this cultural understanding.

We use a river metaphor to convey several key notions. The document collection's time evolution, selected thematic content, and thematic strength are indicated by the river's directed flow, composition, and changing width, respectively. The directed flow from left to right is interpreted as movement through time. In Fig. 2, the river flows from November 1959 to June 1961. The horizontal distance between two points on the river defines a time interval. For example, the time interval represented by the distance between the two vertical dotted lines is almost two months. Like a histogram, ThemeRiver uses variations in width to represent variations in strength or degree of representation. At any point in time, the total vertical distance, or width, of the river indicates the collective strength of the selected themes. The collective theme strength of the river is quite strong in March 1961 (near the right side of the figure) where the river is wide; the collective theme strength is much weaker in June 1961 (the far right of the figure) where the river is narrow.

Colored "currents" flowing within the river represent individual themes. A current's vertical width narrows or broadens to indicate decreases or increases in the strength of the individual theme at any point in time. In Fig. 2, the cyan current represents "weapons"; the weapons theme is relatively weak in November 1959 and relatively strong in December 1959. A current maintains its 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 in the same color and position relative to the other themes. The weapons theme disappears in the months of January and February 1960, but reappears in March 1960.

The entire river may "dry up" for a period of time if the composite currents all disappear. This is possible because a river can effectively represent only a subset of themes from a document collection, typically a few dozen selected themes. This subset may represent only a small portion of the total collection content. We discuss this later in the paper.

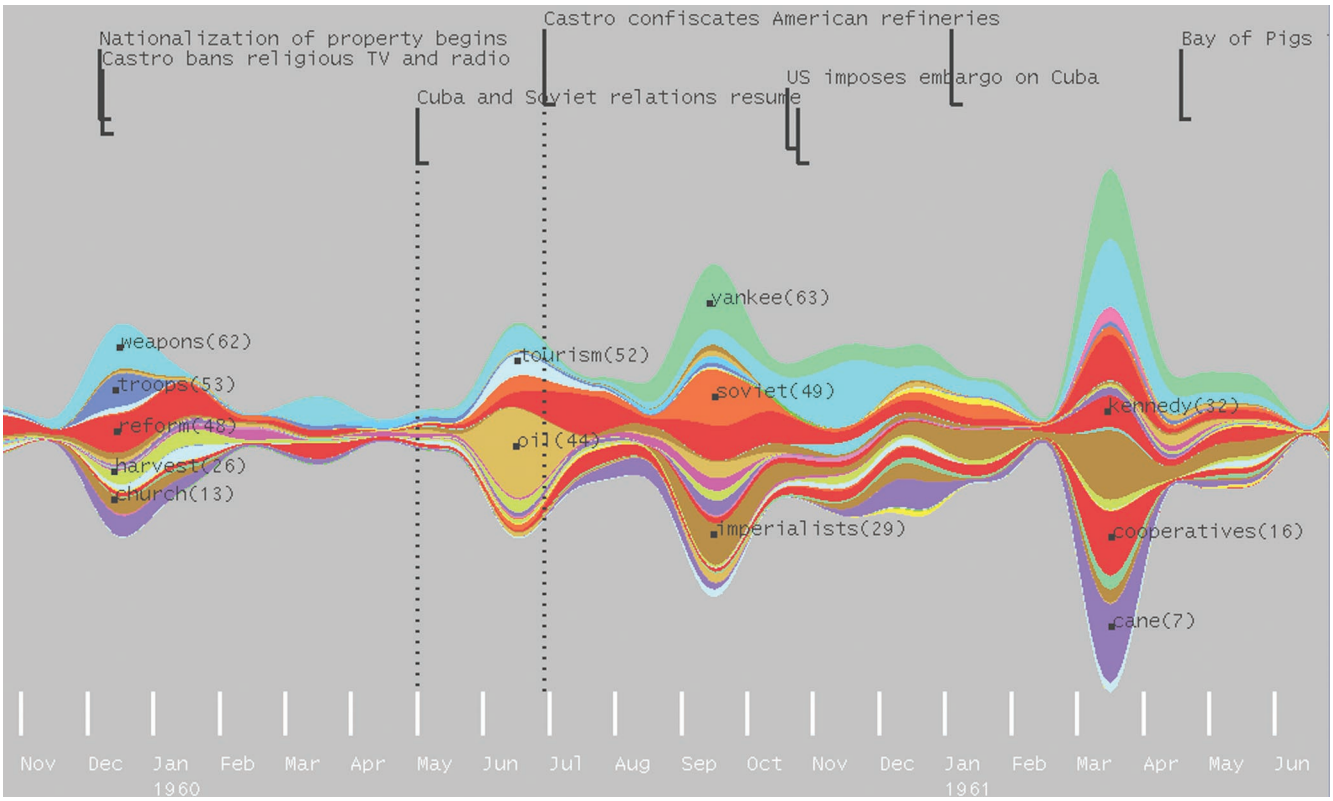

Fig. 2. ThemeRiver uses a river metaphor to represent themes in a collection of Fidel Castro's speeches, interviews, and articles from the end of 1959 to mid-1961.

2.2 Cognitive Considerations

The Gestalt School of Psychology [6], founded in 1919 in Germany, theorized that, for 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 a recent book [7], 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 objects that contain abrupt changes [8].

Smooth, continuous curves bound a theme current in the ThemeRiver visualization. A theme current is assigned a single color for the entire length of the river. The smooth bounds and distinct color help the user track and compare a current's behavior along the river. At a glance, the user can see the pattern of the current as an object—where it bulges, where it shrinks, and where it remains unchanged. We naturally associate the size (area) of the object with strength; a larger area indicates more strength, while a smaller area indicates less strength. The absence of the object indicates no use or strength at that time.

Proximity afforded by stacking the currents makes it easy to compare the current shapes (smoothly bounded within the same time interval or across neighboring intervals. We can see when two patterns match, when they complement each other, or when they appear to be uncorrelated. We can also see when the patterns are concurrent and when they are offset in time. The stacking order of the currents is consistent along the full length of the river. The consistent use of color and stacking order helps the user identify currents as the locus of attention moves across the visualization.

2.3 Context

The ThemeRiver visualization includes the river of theme currents, a time line below the river, and markers for related historical events along the top. Providing such context allows users to evaluate content in relation to issues beyond those contained within the documents themselves. Continuing with the earlier example of candidates running for election, we might ask how the candidates' themes change in response to news events. Do their speeches appear to trigger news events? Does a candidate's opinion have any apparent impact on the stock market?

Fig. 2 shows an interesting example of a related theme and event in the sudden expansion of the "oil" theme just before Castro confiscated American oil refineries in 1960. On occasion, we find patterns than cannot be explained until further investigation uncovers events not included in our event stream. For example, in a later period (not shown), ThemeRiver reveals the use of themes "kennedy" and "missiles" in March 1992. These themes seem outdated for 1992. On further investigation of events in 1992, we discovered that Castro spoke in March at a conference marking the 30th Anniversary of the Cuban Missile Crisis (October 1962). In such cases, we can easily add a marker to

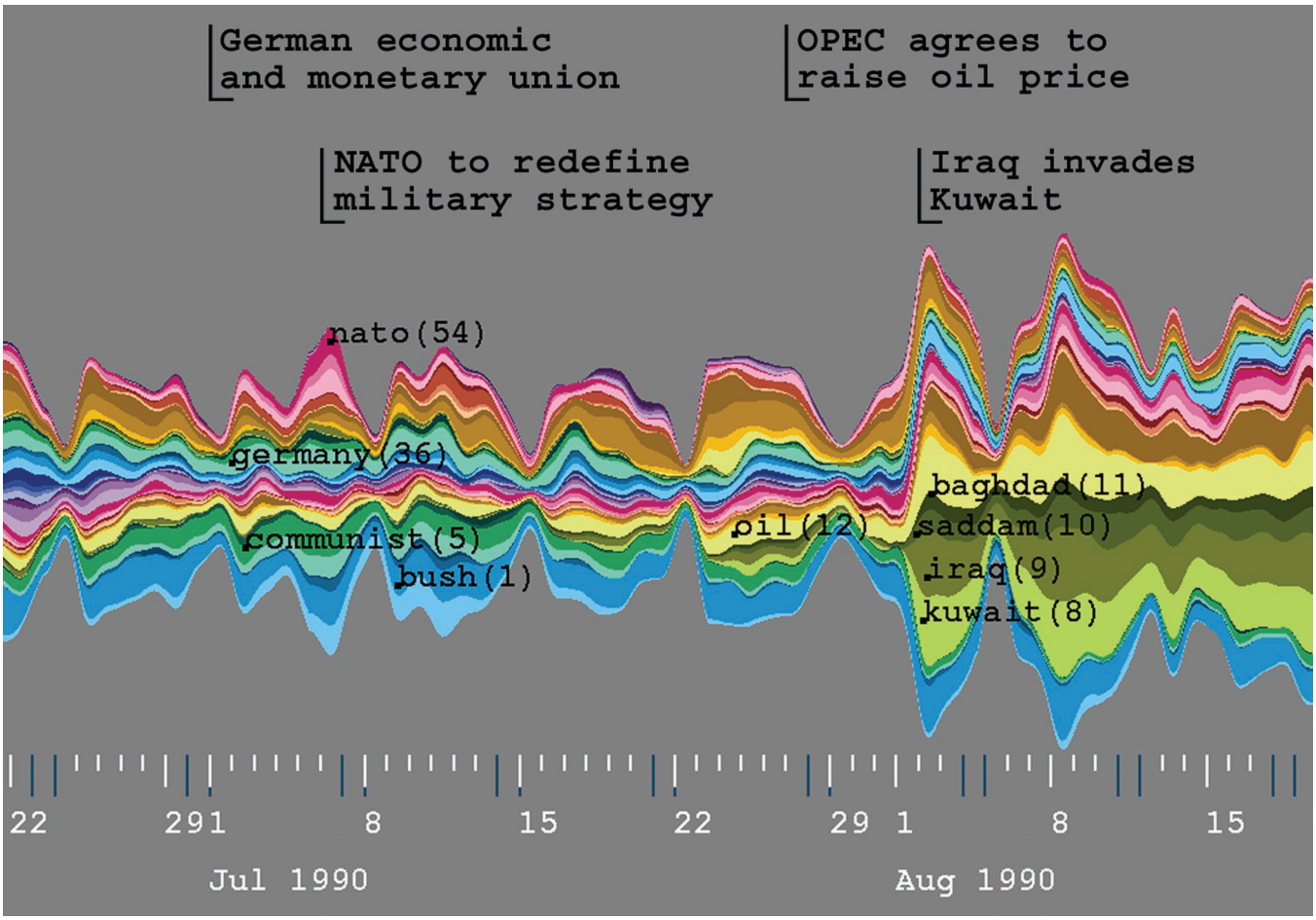

Fig. 3. ThemeRiver visualization of Associated Press news wire stores from July and early August 1990.

the event stream. A user could also add analysis annotations in the same way.

3 Sample Analysis Using ThemeRiver

Hetzler et al. [9] describe the results of a comparative analysis on a single data set using a variety of visual analysis tools. The data set is a collection of over 100,000 Associated Press (AP) news wire stories from 1990 obtained from the TREC3 distribution disks. The comparative analysis focused on large theme changes surrounding the Iraqi invasion of Kuwait on 2 August. Below, we discuss a sample exploration of the same well-studied data set using the ThemeRiver visualization to see if it revealed new information.

To begin, the user selects a set of themes and creates a ThemeRiver visualization of the 1990 AP data. Fig. 3 shows ThemeRiver with the July through early August AP data. The user might begin to explore with a high-level survey of the visualization by panning along the course of the river. The user might look for obvious (large) changes in the pattern, such as the increased total river width in early August. Most of this change can be attributed to the increased strength of a few themes: "kuwait" (in pale green, near the bottom of the river), "iraq" (darker green, above "kuwait"), "saddam" (still darker green, above "iraq"), "baghdad" (the darkest green, near the center of the river), and "oil" (pale yellow, also near the center of the river). These theme currents correspond to the 2 August 1990 Iraqi invasion of Kuwait.

The user might look for persistent themes such as the yellow current running through the center of the river representing "oil." The user might also look for narrow currents in the river that signal relatively light use of particular themes, such as the narrow band of magenta at the top representing "nato."

The invasion stories were explored in the earlier analysis with other visualization tools. However, ThemeRiver reveals some additional detail not noted in the previous study. The theme "oil" is persistent across the image, indicating that oil was a relatively common topic throughout the period, though talk about oil increased markedly during the invasion. In the five weeks before the invasion, the themes of "iraq," "saddam," and "baghdad" appear briefly in relatively narrow currents. The themes "kuwait" and "saudi" (brown, above "oil") show up in bursts the two weeks before the invasion. News stories corresponding with these topical bursts covered the verbal conflicts leading up to the invasion. It is clear from ThemeRiver that these themes gained strength over the weeks preceding the invasion, gaining most strongly in the week immediately prior to the invasion. These observations are not possible with other visualizations and demonstrate two advantages. First, the theme current patterns allow a user to

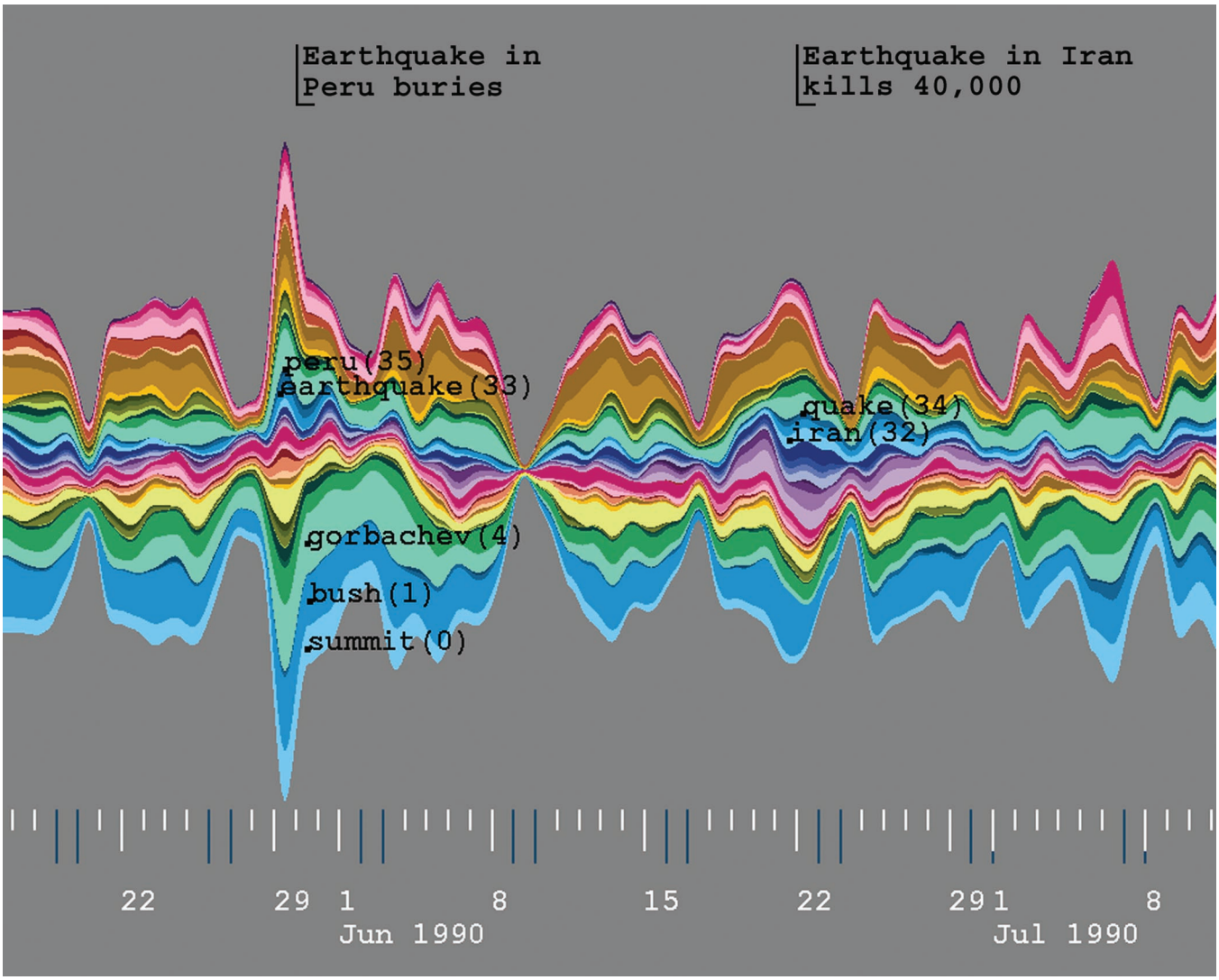

Fig. 4. ThemeRiver of AP data from June-July 1990 identifies very different events from those revealed immediately afterward (Fig. 3).

easily distinguish between persistent and bursty themes. Second, a user can observe subtle changes, such as the building news coverage of the oil producing countries in the Middle East.

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

Fig. 4 shows the ThemeRiver from earlier in the summer of 1990. On 29 May, a large change in theme strength appears that does not match any previously identified events. Some of the larger currents here are "gorbachev" (light green near the bottom of the river), followed by "bush" (turquoise), and "summit" (pale blue). Viewing the pertinent news documents from that time, we find that several world leaders, including Bush and Gorbachev, participated in a four-day summit meeting in Washington. On the same peak day, an earthquake occurred in Peru. ThemeRiver helps identify coocurrences and patterns in time; the analyst would explore further to decide whether such events are connected.

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 (Figs. 3 and 4), the river seems to narrow quite frequently. On closer inspection, we see that the narrow sections correspond to Sundays. Because the river contains only a subset of the themes in the collection, at this point we cannot tell from the ThemeRiver 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 usability evaluation. In response, we have added an optional histogram representing the total number of documents for each time slot, along with the portion represented by the themes in the river. In Fig. 5, the white regions of the histogram represent the documents containing themes included in the ThemeRiver currents, while the dark gray represents the remainder of documents

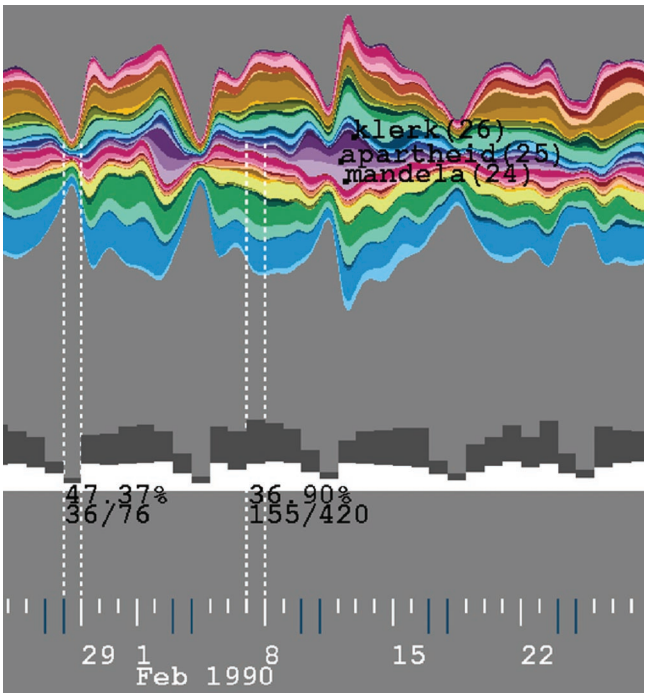

Fig. 5. The addition of a histogram to ThemeRiver reveals that news is light on Sundays, not that themes shift.

in the collection for that day. A user can click on a histogram bar to see the percentage as well as the ratio of the number of documents with our theme subset to the total number of documents. With this histogram, it is apparent that there are generally fewer AP news stories released on Sunday than other days.

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. Fig. 6 shows two such parallel rivers: The upper river shows the AP news stories from New York while the lower river shows news stories from Washington, D.C., for the same time period. Some differences in major themes are immediately apparent. The Washington themes emphasize "bush" (turquoise, bottom), "supreme" (dark brown at the top, refers to the Supreme Court), and "senate" (tan current under "supreme"). The New York stories show a major growth in the themes "apartheid"(purple, near top) and "mandela" (lavender, under "apartheid"), corresponding with Nelson Mandela's visit to the US the week ending 22 June. He arrived first in New York, where he spent several days before proceeding to Washington, D.C. Relatively narrow currents for "apartheid" and "mandela" appear the following week in purple and lavender near the center of the lower river.

4 Implementation Details

We have implemented a proof-of-principle prototype of ThemeRiver and used it to explore data from multiple sources. This section describes some of the design work required to implement the ThemeRiver prototype.

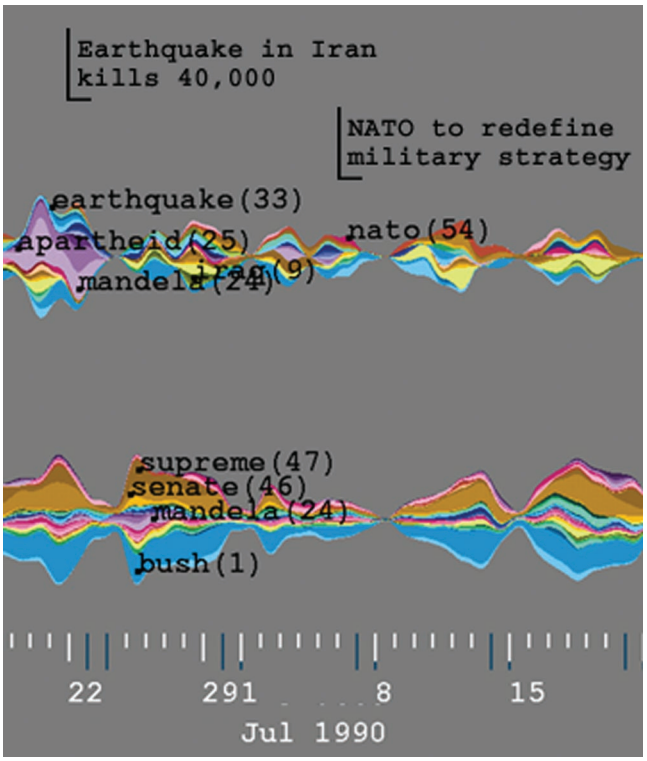

Fig. 6. Parallel rivers let users compare AP data from Washington, D.C. (bottom river) and New York (top river) from the same time period.

4.1 River Data

Time-tagged electronic documents are easily binned into appropriately sized time intervals. Our initial data set is a collection of speeches, interviews, articles, and other text associated with Fidel Castro spanning a 40-year period. We separate these documents into one-month intervals based on their date tags.

Next, we calculate the strength of each theme in each interval. The strength can be calculated in various ways. For example, the number of an interval's documents containing the theme word could represent the strength of that theme in that interval. Alternately, the number of occurrences of a theme word in an interval's documents could also represent the strength.

For our test data set, we focus on a set of 64 theme words, a subset of topic word candidates automatically calculated from the Castro document collection. For each of these theme words, we counted the number of documents in each time interval containing the theme word. The counts y_{ji} of the speeches containing the jth theme word during the ith time interval, and the associated time intervals t_i, formed the input data for the ThemeRiver representations of the Castro collection.

4.2 External Event Data

For the prototype, we manually create a simple ASCII file of external events for each data set. The events are listed in time order. Each event occupies one line. New events or annotations are easily added to the file.

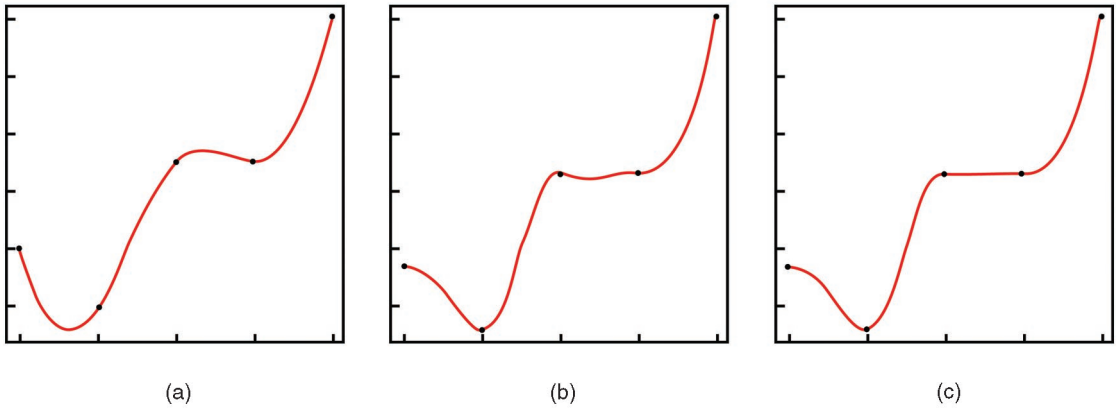

Fig. 7. (a) Interpolating cubic spline. (b) Interpolating spline augmented with derivative data. (c) Interpolating spline augmented with derivative data and enforced constraint.

4.3 River Calculations

This section describes an algorithm for creating the smooth curves shown in the ThemeRiver display. First, we describe the data, then desired characteristics of the algorithm, and, finally, the algorithm used.

4.3.1 Data Details

The data are collections of pairs, where the first coordinate is often interpreted as time and the second is an associated, nonnegative magnitude. For simplicity, we assume that the first coordinate values are the same for all the pairs. So, we can denote the data as: (t_i, y_{ji}) where $i = 1, \ldots, n$ and $j = 1, \ldots, m$. We have $t_i < t_{i+1}$ and $y_{ji} \geq 0$. For the Castro collection example in this paper, t_i indicates successive months and y_{ji} is a measure of to what extent documents in the t_ith month refer to the jth topic. The ThemeRiver visualization shows a smoothly varying presentation of these data in which the values and the associated smooths are centered and stacked. The width of the jth band at time t_i indicates the interpolated value of y_{ji} at that time.

4.3.2 Algorithm Properties

The goal is to create aesthetically pleasing curves that fit the data; for simplicity, we temporarily drop the first subscript from y. Consider data (t_i, y_i), where $t_i < t_{i+1}$ and $y_i \geq 0$. Denote the curves by $\hat{f}$, the desired properties of the interpolator include:

Property 1. $\hat{f}(t_i) = y_i$, that is, $\hat{f}$ is an interpolator. While smoothing is often used to reduce the noise in data, we decided to focus on constructing a display that was faithful to its inputs. If some sort of smoothing or regression is desired, then the information shown in ThemeRiver could be the smoothed version of the data.

Property 2. $\hat{f}(t)$ is between $\hat{f}(t_i)$ and $\hat{f}(t_{i+1})$ for $t \in [t_i, t_{i+1}]$. Linear interpolators satisfy these conditions, but natural cubic spline interpolators often do not [10]. For nonnegative data, Property 2 implies $\hat{f} \geq 0$. This property is needed so that the stream widths stay positive, allowing for better visual tracking of streams across time.

An interpolator satisfying these conditions can be stacked to obtain the ThemeRiver view as follows:

Step 1. Calculate the interpolator at a mesh of points spanning the duration of interest. Denote the mesh coordinates by s_i and the corresponding interpolant values by $\hat{f}_{ji}$ (we've resumed with the first subscript).

Step 2. Calculate the stack sum at each s_i: $f_{\bullet i} = \frac{1}{2}\sum_{j=1}^{m} f_{ji}$ and create the centered, cumulative values $\tilde{f}_{ji} = \sum_{j'=1}^{j} f_{j'i} - f_{\bullet i}$; also let $\tilde{f}_{0i} = -f_{\bullet i}$.

These curves are plotted and the colored bands in the ThemeRiver figures are filled between these curves. Thus, the height of the band between $\tilde{f}_{ji}$ and $\tilde{f}_{j+1,i}$ has the natural interpretation as an estimate (or value) of the curve corresponding with the ordinate s_i.

4.3.3 Algorithm

Finally, we describe the construction of the interpolator. The mathematical properties of this function are outlined above, and the aesthetic goal is to achieve a good-looking curve. Fig. 7 shows the evolution of the curves in ThemeRiver. The data being interpolated are shown as the solid dots. The first of these, Fig. 7a, is a standard interpolating cubic spline. For these data, the interpolant fails to satisfy Property 2. We observe that, in order for the interpolant to satisfy Property 2 at a local extreme, the derivative of the curve must be zero at that point. Thus, a preprocessing step became finding points t_i at which y_i is a local, nonstrict, extrema. For the data shown in Fig. 7b, the middle three points are relative extrema (minimum, maximum, minimum, respectively). At these points, we augment the data with a zero derivative condition.

A mathematical framework that readily enables calculating the curve with this additional information is presented in Wahba [11], especially chapters 1 and 2. Following her notation, we're looking for a curve in W_m that satisfies the constraints:

Constraint 1. Interpolation: $f(t_i) = y_i$.

Constraint 2. Zero derivative at constraint points: $f'(t_{i'}) = 0$.

Both of the above types of constraints are point evaluations of a continuous linear functional, the first type being evaluation at a point and the second being evaluation of a derivative at a point. In the framework presented in

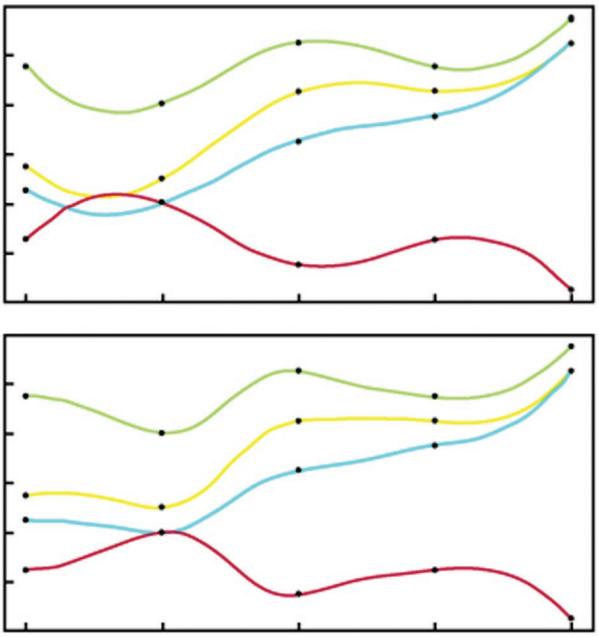

Fig. 8. A comparison of two images demonstrating the effect of Property 2 on stacked curves. In the top image, the red curve overplots two others. The bottom image shows the result when the curves are constrained to satisfy Property 2.

Wahba [11], the minimum W_m-norm interpolant can be readily found based on standard matrix decompositions and operations. Note that the standard interpolating spline of order m is obtained as a minimum W_m-norm function satisfying the first constraints. We simply add additional interpolation information to help achieve the desired properties. Fig. 7b shows the result of applying these additional constraints to our data.

The resulting function is an improvement with respect to the aesthetic goal; however, the interpolant is not constant between the identical successive points. We address this characteristic by a postprocessing step that replaces the interpolant with values that satisfy the aesthetic goal, when necessary:

- For $s_{i'}$ between t_i and t_{i+1}; if $\hat{f}(s_{i'}) < y_i$, replace $\hat{f}(s_{i'})$ with $\min(y_i, y_{i+1})$. If $\hat{f}(s_{i'}) > y_i$ replace $\hat{f}(s_{i'})$ with $\max(y_i, y_{i+1})$.

This final adjustment creates the curve in Fig. 7c. These are the smooth, interpolating curves that are stacked to create the ThemeRiver views presented in this paper. Fig. 8 shows the effect when the curves are stacked in the river.

Fig. 9 illustrates how various design goals and corresponding curve creation algorithms affect the resulting river portrayal. Fig. 9a shows an algorithm based on smoothing (as opposed to interpolating) the data; it also does not satisfy the shape-preserving goal. Users who saw actual data points plotted with respect to such curves were concerned that the curves "did not really represent the data." In addition, this algorithm occasionally led to a discontinuity in the derivative, such as the kink in the lower curves in mid-July. Such discontinuities draw the eye to a particular point in the display, an undesirable effect when it's due solely to the portrayal and not to the data.

The curves in Fig. 9b interpolate the data points and avoid the derivative discontinuities but violate the aesthetic goal. The river portion for the beginning of August shows several curves that overshoot the relative maxima; as a result, they are obscured by other curves.

The river shown in Fig. 9c is created with the final algorithm as described above.

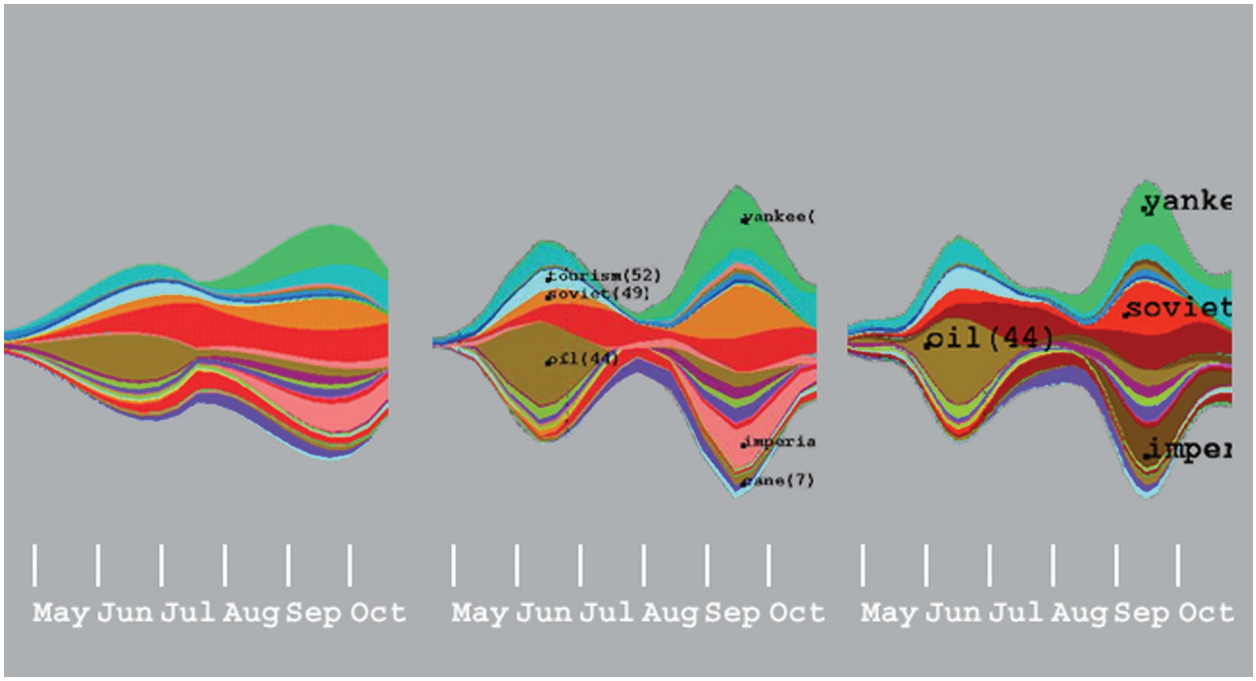

Fig. 9. The section of ThemeRiver on the left, (a), shows the algorithm based on smoothing. The section in the center, (b), shows the algorithm with interpolation but violates Property 2. The section on the right, (c), shows the outcome of the final algorithm.

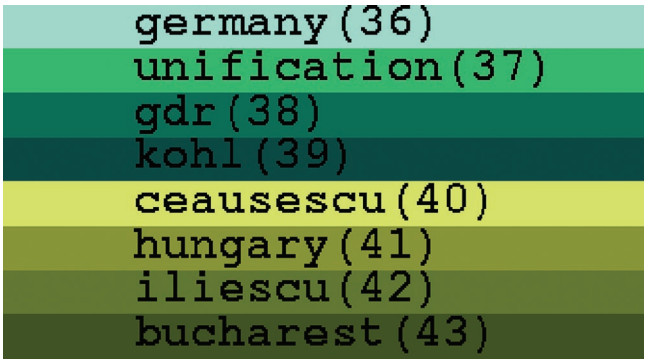

Fig. 10. Tracking related themes is simplified by assigning them to the same color family. This ensures that related themes appear together and are identifiable as a group.

4.4 River Colors

Color choices for ThemeRiver 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 between colors. Currently, we are exploring a solution suggested during formative usability evaluation: sorting themes into related groups and displaying each group with a color family. Fig. 10 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.

5 Usability Evaluation

Early in developing ThemeRiver, 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?

We used 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 Fig. 11). We made the content of the histogram as similar as possible to ThemeRiver's content.

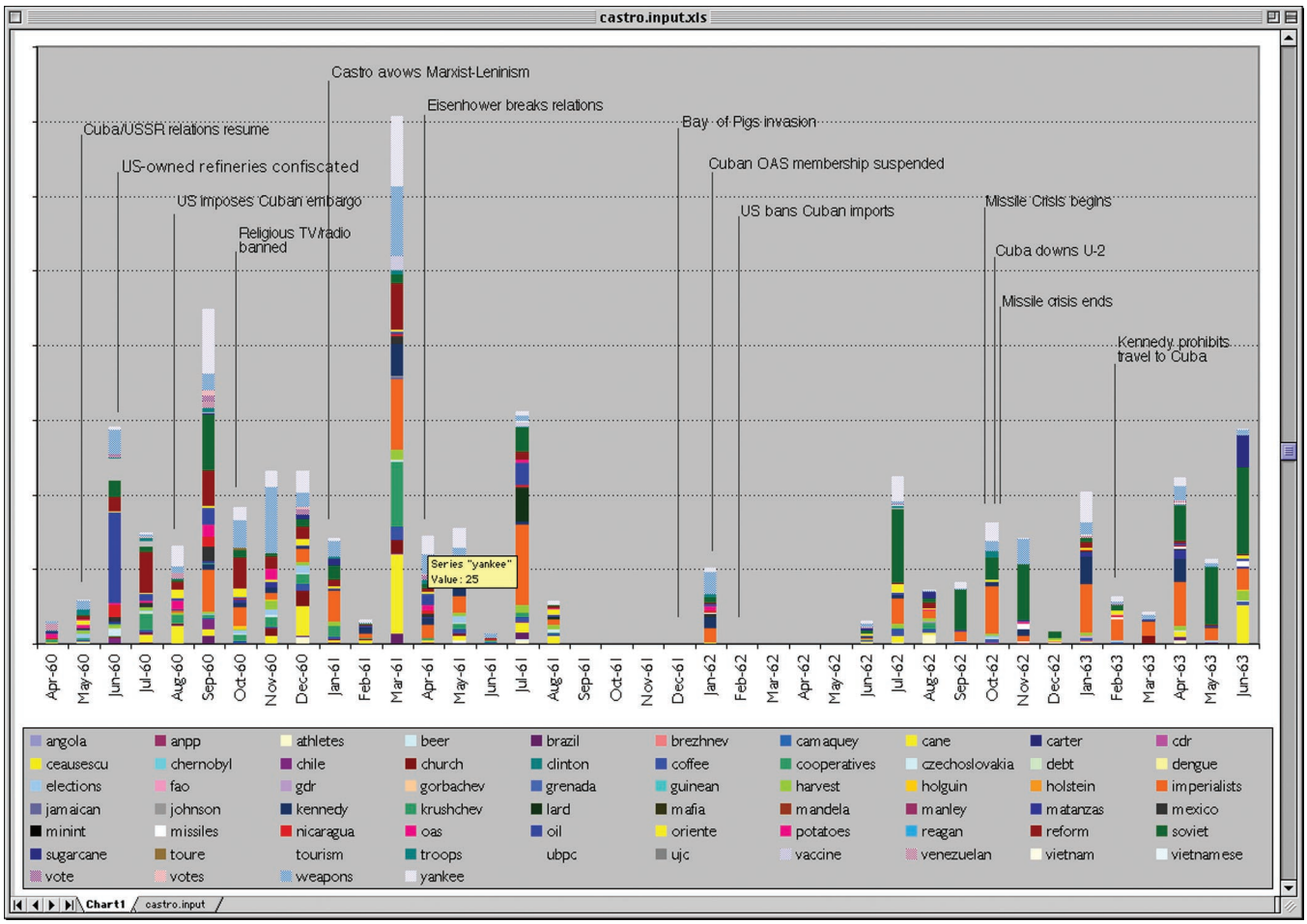

Fig. 11. Like the ThemeRiver in Fig. 2, this histogram uses the Castro collection data and depicts changes in thematic content over time.

For example, the histogram depicted thematic content by months, using the same values that drive ThemeRiver. The month time line 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 histogram and the other one began with ThemeRiver.

We asked each participant questions about what they observed in each display. Some specific questions included:

- In July 1962, what are the three most discussed themes?
- Where is a new theme introduced?

Some examples of more general questions that we asked were:

- 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 also asked participants to complete a short questionnaire, eliciting 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 currents 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 further.

Questionnaire responses showed that users found ThemeRiver easy to understand. They 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 deemphasize very small values. We asked about the value of the river metaphor and users rated it high as well. They observed that the connectedness of the river helped them follow a trend more easily over time than they were able to do using the histogram. This result is compatible with the perception principles described by Ware [8], particularly the value of connectedness and smooth curvature.

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 the participants liked ThemeRiver's abstraction to the whole collection and thus away from individual documents, both of them suggested adding features to access documents on demand. 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 also wanted to select a current and see the documents that contributed to it. And, the participants wanted to be able 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.

Many of the user suggestions drove the design improvements noted earlier in this paper, such as color families and the addition of an optional histogram to show document representation in the river. Others remain as future work.

6 Design Challenges

Focusing on themes rather than documents raises 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 that contain 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 Section 4.5, or using each current to represent a set of themes rather than a single theme.

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. However, 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. But, 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 address 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 Related Work

Many information visualization systems include features for viewing time changes within an existing portrayal of

other features. There are also a few that focus more directly on time portrayals. We review both categories briefly.

One method for including time within an information visualization is to filter what is shown by the selected time period and then update the display as the selected time period changes. Systems like FilmFinder use this method, enabling the user to dynamically change the time window of interest and immediately see the effect on the displayed results. This approach has been called dynamic queries [12].

Another common approach is to show time as an attribute of the information objects. Virginia Tech's Envision system lets users map document metadata values, such as date, type, size, and relevance to a query, to a selection of graphical encodings, such as location along the x-axis or y-axis, color, shape, or size [13]. The goal is to give users full control over the mapping to create the encodings that work best for the user and task at hand.

The SPIRE Galaxies visualization [1] includes mapping of document time to line angle and length, designed to highlight the most recent documents while allowing comparison relative to other time periods. Icons representing documents are placed in a 2D plane, with location representing similarity among the documents' themes. The user selects a set of contiguous discrete time periods, such as several days. Documents from the most current day are shown as long vertical lines. Older documents' icons use line angles progressively varying between vertical and horizontal and shortening in length [14].

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 [15], [16]. The visualization displays time along the x-axis and uses the y-axis to categorize events. Horizontal bars depict duration for events such as hospitalizations, while tick marks are used for discrete events such as medical tests. Graphical attributes such as color and width can show event relationships or importance. TmViewer uses a similar approach, adding the ability to show parent-child relationships with lines between related time bars [17].

The DIVA system [18] uses animation to show how particular measured features change in relation to the temporal flow of a video. The video plays in the center of a window. The measured features are shown as colored rectangular marks along the top and left edges of the window; as the video plays, the presence and color of the rectangles change to show feature values corresponding to the video sequence. To add more time context, the "past" values do not simply disappear. It's as if each feature has a spatial channel in which it is portrayed, roughly corresponding to where the feature would appear in a series of cascading windows. As the video progresses, the feature values seem to approach from the lower and right directions, jump to the planar window containing the video, and then recede toward the top and left.

To help groups collaborating to create a document or other artifact, the Timewarp system developed at Xerox PARC [19] lets users view and edit multiple time lines of the changing state of that artifact. The metaphor used is similar to a state diagram, with lines connecting state nodes and branches.

Karam [20] describes a time line display generator, a prototype that can be configured to interpret and display events from a system under analysis. As an example, it was used to display event data from a desktop videoconferencing system, to identify perceived problems with that system.

Tufte [21] presents an artist's illustration showing trends in music, which is similar in concept to the ThemeRiver. 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.

8 Conclusions and Future Work

ThemeRiver is a demonstration prototype developed to test the value of the river metaphor for revealing patterns, relationships, and trends in themes of a document collection. The response by users and demonstration viewers has been very positive. The metaphor is intuitive and the stacked, smoothly bounded areas of themes are easily observed and compared. Usability testing and demonstration viewer comments confirm the value of providing context information. We conclude that ThemeRiver is a potentially valuable tool for information analysts and plan to develop it into a full visualization tool.

Nevertheless, there remain some important limitations to overcome; the most pressing are faster calculation and rendering to support interesting user interactions. Plans for future work include the necessary speedup work, increased user control of theme selection and layout, the capability to drill down to the underlying documents, and additional contextual support such as automatic event marker generation.

Acknowledgments

The authors gratefully acknowledge the contributions of our colleagues at Battelle for the development and testing of the ThemeRiver prototype. Special thanks for contributions to this paper go to Grant Nakamura, Alan Willse, Sharon Johnson, Sharon Eaton, and Wanda Mar. Battelle Memorial Institute's Information Technology Platform funded this research.

References

[1] 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," *Readings in Information Visualization: Using Vision to Think,* S.K Card, J.D. Mackinlay, and B. Shneiderman, eds., pp. 442-45, San Francisco: Morgan Kaufmann, 1999.

[2] D. Brodbeck, M. Chalmers, A. Lunzer, and P. Cotture, "Domesticating Bead: Adapting an Information Visualization System to a Financial Institution," *Proc. InfoVis '97,* pp. 73-80, 1997.

[3] X. Lin, "Map Displays for Information Retrieval," *J. Am. Soc. Information Science,* vol. 48, no. 1, pp. 40-54, 1997.

[4] E.R. Tufte, *The Visual Display of Quantitative Information,* p. 28. Cheshire, Conn.: Graphics Press, 1983.

[5] G. Lakoff and M. Johnson, *Metaphors We Live By.* Chicago: Univ. of Chicago Press, 1983.
[6] K. Koffka, *Principles of Gestalt Psychology.* New York: Harcourt-Brace, 1935. C. Ahlberg, and B. Shneiderman, "Visual Information Seeking: Tight Coupling of Dynamic Query Filters with Starfield Displays," *Proc. Conf. Human Factors and Computing Systems (CHI '94),* pp. 313-317, 479-480, 1994.
[7] D.D. Hoffman, *Visual Intelligence: How We Create What We See.* New York: W.W. Norton & Company, 1998.
[8] C. Ware, *Information Visualization: Perception for Design.* San Diego, Calif.: Academic Press, 2000.
[9] B. Hetzler, P. Whitney, L. Martucci, and J. Thomas, "Multi-Faceted Insight through Interoperable Visual Information Analysis Paradigms," *Proc. IEEE Symp. Information Visualization (InfoVis '98),* pp. 137-144, 1998.
[10] G. Wolberg and I. Alfy, "Monotonic Cubic Spline Interpolation," *Proc. Computer Graphics Int'l '99,* June 1999.
[11] G. Wahba, *Spline Models for Observational Data.* Philadelphia: SIAM, 1990.
[12] C. Ahlberg and B. Shneiderman, "Visual Information Seeking: Tight Coupling of Dynamic Query Filters with Starfield Displays," *Proc. Conf. Human Factors and Computer Systems (CHI '94),* pp. 313-317, 479-480, 1994.
[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," *Proc. ACM Conf. Research and Development in Information Retrieval (SIGIR '96),* pp. 67-75, 1996.
[14] L.T. Nowell, E.G. Hetzler, and T. Tanasse, "Change Blindness in Information Visualization," in press.
[15] C. Plaisant, D. Heller, J. Li, B. Shneiderman, R.J. Mushinlin, and J. Karat, "Visualizing Medical Records with LifeLines," *Conf. Human Factors and Computing Systems (CHI '98) Summary,* pp. 28-29, 1998.
[16] C. Plaisant, B. Milash, A. Rose, S. Widoff, and B. Shneiderman, "Lifelines: Visualizing Personal Histories," *Proc. Conf. Human Factors and Computing Systems (CHI '96),* pp. 221-227, 1996.
[17] V. Kumar and R. Furuta, "Visualization of Relationships," *Proc. Hypertext '99,* 1999.
[18] W. Mackay and M. Beaudouin-Lafon, "Diva: Exploratory Data Analysis with Multimedia Streams," *Proc. Conf. Human Factors and Computing Systems (CHI '98),* pp. 416-423, 1998.
[19] K.W. Edwards and E.D. Mynatt, "Timewarp: Techniques for Autonomous Collaboration," *Proc. Conf. Human Factors and Computing Systems (CHI '97),* pp. 218-225, 1997.
[20] G.M. Karam, "Visualization Using Timelines," *Proc. 1994 Int'l Symp. Software Testing and Analysis,* pp. 125-137, 1994.
[21] E.R. Tufte, *Visual Explanations: Images and Quantities, Evidence and Narrative,* pp. 90-91. Cheshire, Conn.: Graphics Press, 1997.

Susan Havre received the MS degree in computer science from Washington State University in Pullman, Washington. She is a senior research scientist in the Synthesis, Analysis, and Visualization of Information Group at Battelle Pacific Northwest in Richland, Washington. Her research interests include the visualization and mining of large-scale data, both scientific and textural. She is a member of the IEEE Computer Society and the ACM.

Elizabeth Hetzler received the BA degree in mathematics and French from Vanderbilt University in Nashville, Tennessee, and the MS degree in computer science from Washington University in St. Louis, Missouri. She is a chief scientist in the Synthesis, Analysis, and Visualization of Information Group at Battelle Pacific Northwest. Before coming to Battelle, she worked at McDonnell Douglas, performing research in CAD/CAM and human-computer interaction. Her current research interests include information visualization, text and data mining, and human information interaction.

Paul Whitney received the BS degree in mathematics from Oklahoma University and the PhD degree in statistics from the University of Wisconsin Madison. He has worked at Battelle Pacific Northwest for 10 years and is currently a staff scientist in the Statistical and Quantitative Sciences Group. Before coming to Battelle, he was an assistant professor of statistics at the University of Texas at Dallas and Southern Methodist University. His current research interests include exploratory analysis of mixed-type, multivariate data, text analysis, image analysis, state-space modeling and Kalman filtering, event detection, and forecasting.

Lucy Nowell holds the MA degree in theatre (design), MFA degree in drama (design), and the MS degree in computer science. She received the PhD degree in computer science from Virginia Tech. She is a chief scientist in the Synthesis, Analysis, and Visualization of Information Group at Battelle Pacific Northwest. Her interests include information visualization, dynamic user modeling and intelligent interfaces, multimodal user interfaces, ubiquitous computing, virtual and augmented realities, digital electronic libraries, methods of formative usability evaluation, and user interface design.