ThemeRiver™[*]: In Search of Trends, Patterns, and Relationships

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

ABSTRACT

ThemeRiver™ is a prototype system that visualizes thematic variations over time across a collection of documents. The "river" flows through time, changing width to depict changes in the thematic strength of documents temporally collocated. Themes or topics are represented as colored "currents" flowing within the river that narrow or widen to indicate decreases or increases in the strength of a topic in associated documents at a specific point in time. The river is shown within the context of a timeline and a corresponding textual presentation of external events.

Keywords

visualization, visualization metaphors, trend analysis, timeline

1. INTRODUCTION

ThemeRiver is designed to facilitate the identification of trends, patterns, and unexpected occurrence or non-occurrence of themes or topics. In particular, ThemeRiver provides users a macro-view of thematic changes in a corpus of documents over a serial dimension. In our prototype, we use time as the serial dimension. We provide contextual information through a timeline and give markers for co-occurring events of interest. Users interact with the visualization to explore the information space and to discover trends, patterns, and other features of interest.

Many information visualization systems represent each document or group of documents with a glyph or icon, based on various document attributes. Few systems represent changes over time at all, but in those that do, changes over time are typically represented by a sequence of images or "frames." Each frame displays only the documents related to a specific time slot. For example, in SPIRE's Galaxies visualization [8], the Time Slicer Tool can be used to view only the icons of documents originating within a specified time period. A user may choose to see a corpus that spans a ten-year period partitioned into ten images/frames, displaying only the documents associated with each year; the tool can then sequence through these frames in temporal order. This type of visualization is document-centric.

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

A histogram might provide one approach to theme-centric visualizations, with each bar representing a time slice, and color variations within the bar representing the relative strength of themes specific to the time slice. However, understanding the histogram requires the user to work at relating the themes across time, because the bars are always 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. (See Figure 1.) However, ThemeRiver simplifies the user's task of tracking individual themes through time by providing a continuous "flow" from one time point to the next. The horizontal flow of the river represents the flow of time. Each vertical section of the river corresponds to an ordered time slice. Each theme is represented by a colored current that runs horizontally within the river. The width of a current changes to reflect the strength of the corresponding theme over time. As the occurrence of a theme increases over time, the corresponding current widens. As a theme's occurrence decreases over time, the corresponding current narrows. 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 reoccurs, then the current disappears and then reappears with the same color representation and in the same position relative to other themes. We believe this metaphor is familiar and easy to understand and that it requires little cognitive effort to interpret the visualization.

[*] Patent pending.

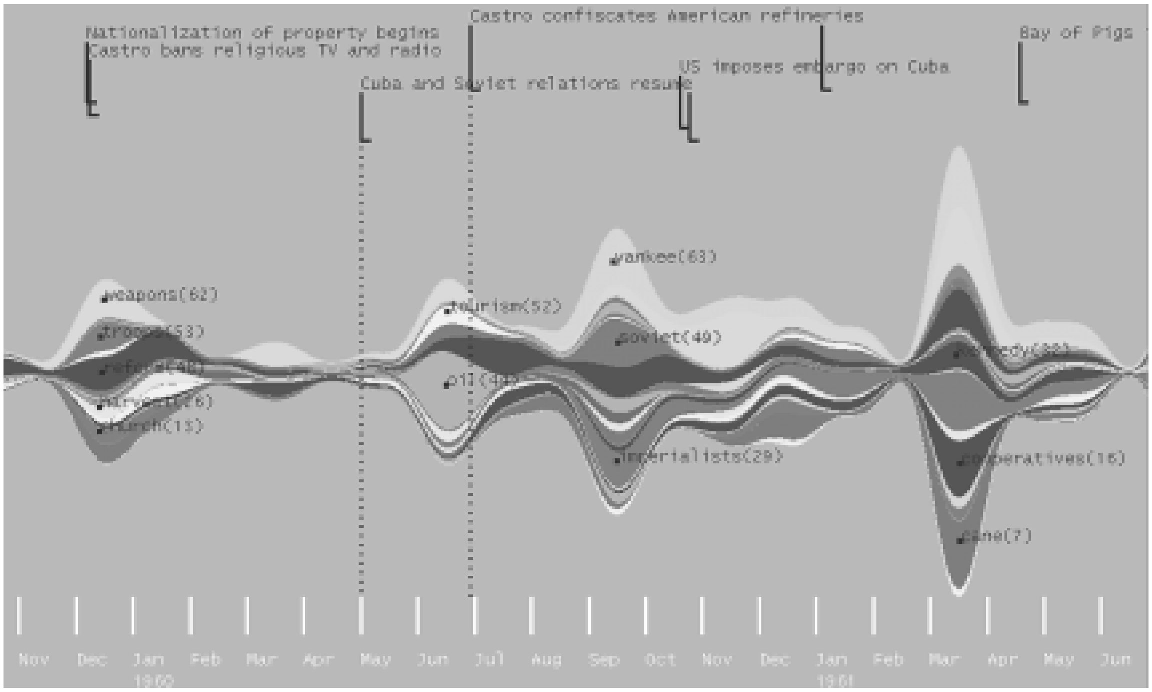

Figure 1: ThemeRiver showing Castro data from November 1959 through June 1961.

2. RELATED WORK

We know of no other systems that use the river metaphor to depict the passage of time. However, a similar idea is presented by Tufte [7] in an illustration showing trends in music. Much research has been done on the representation of time in systems serving various purposes, but space allows mention of only a few examples. Recent work on representation of time has been done for Time Warp at Xerox PARC [1] and for the LifeLines medical records system, developed jointly by the University of Maryland and IBM [6]. Time has been represented in the context of multimedia streaming in Diva [4]. The Envision system at Virginia Tech allows users to explore trends in digital library metadata, including publication dates [5]. Earlier work on use of timelines includes that by Karam [2] and Kullberg [3].

3. VISUALIZATION DESIGN AND INTERACTIONS

We have implemented a proof-of-principle prototype of ThemeRiver and demonstrated it with a collection of speeches, interviews, articles, and other text associated with Fidel Castro over a 40-year period. Values for theme strength across the collection are those produced by the text analysis engine for SPIRE [8]. The visualization, shown in Figure 1, includes the river, a timeline below the river, and markers for related historical events along the top. Users may interact with the visualization to display or hide topic and event labels, to hide or display time and event grid lines, to display or hide the raw data points, and to choose alternate algorithms for line drawing for the currents and river. The user may also display the associated time or topic 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.

The Castro collection invites speculation about the relationship between his words and actions and between the topics he discussed and contemporaneous events. In particular, we are interested in whether changes in themes Castro discussed let users predict subsequent political, economic, or military actions, as well as how he responded to various international events.

To explore selected themes in the Castro collection using ThemeRiver, a user might begin with a high-level survey of the visualization by panning along the course of the river. The user could look for gaps in the river that signal an absence of particular themes. The user might also look for peaks that signal heavy use of the topics, or changes in the colors of the river currents signaling changes in themes.

The event line in Figure 1 shows that in May 1960, Cuba and the Soviet Union resumed diplomatic relations. We see that Castro started talking about the Soviet Union after the resumption was announced. During the 18 months prior to the announcement, he mentioned the Soviet Union only briefly in February 1960. In May and June 1960, we observe that Castro started talking a great deal about oil. At the end of June 1960, he confiscated the oil refineries located in Cuba that were owned by United States companies.

Figure 2 suggests that during the period of July to November 1968, Castro was concerned with political issues as reflected in the prevalence of such topics as Yankee, imperialist, Soviet, and Czechoslovakia. By January 1969, just before he began rationing sugar, Castro switched his focus to agricultural topics, especially sugar and sugarcane. By May he was still talking about sugarcane, but he also began talking about the conflict in Vietnam.

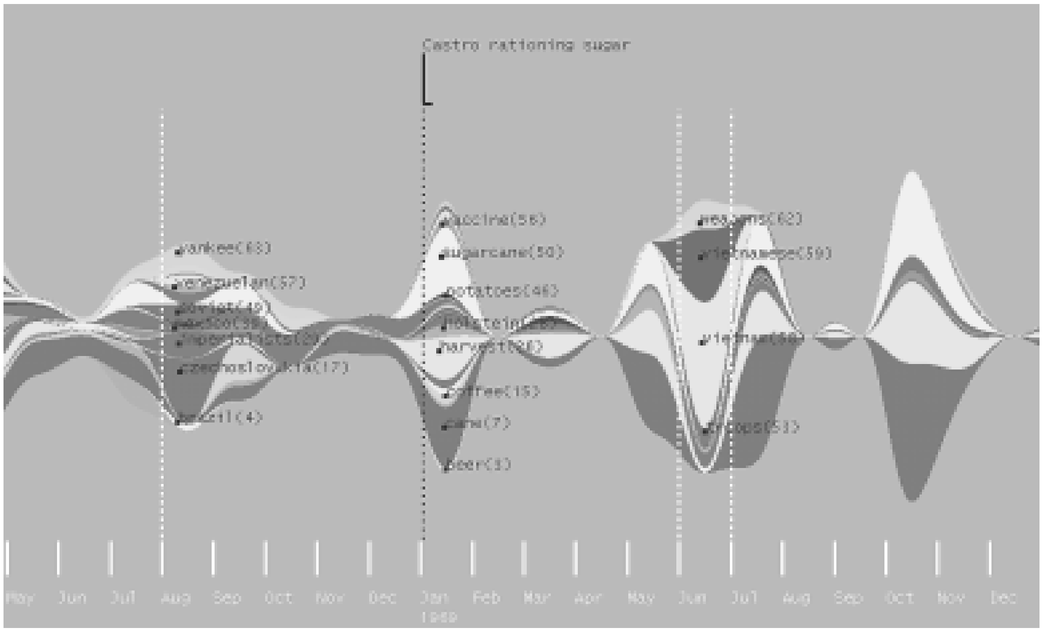

Figure 2: ThemeRiver showing Castro data from May 1968-December 1969.

Figure 3 shows not only theme changes but also unusually heavy emphasis on some new themes. Scanning theme labels and deciding that the theme changes are indeed interesting, the user clicks on currents to toggle more theme labels on and off. The user observes that in March 1992, Castro's main topics are weapons, Soviet, missiles, and Kennedy. An event marker notes that a conference on the Cuban Missile Crisis was held in Havana in March 1992. The user determines that the conference pertains to these topics.

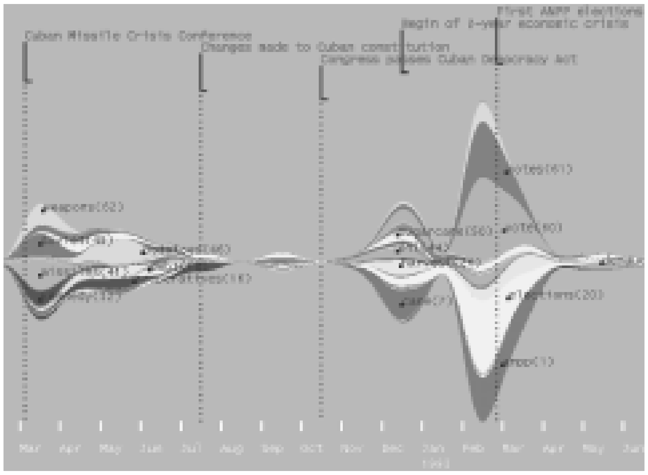

Figure 3: ThemeRiver showing theme changes in the 1990s.

The user sees that agricultural topics dominated in December. In February, Castro started talking a great deal about votes, elections, and ANPP (National Assembly of the People's Government). An event marker shows that the first democratic election in Cuba was scheduled for February. However, in July, the Cuban constitution was changed, and in October the Cuban Congress passed the Cuban Democracy Act. Our user notes that Castro did not have much at all to say during this period. The user is also interested in Castro's silence on the election topics between October and February. The user can look for more external data — maybe Castro did not give any speeches during that period, or perhaps some important documents are missing from the collection.

4. DESIGN CHALLENGES

Ideally, a visual metaphor facilitates discovery by presenting data in an intuitive, easy way that lets users interpret the presentation and data without undue strain. Further, the visualization should not mislead the user.

A key advantage of the river metaphor over a simple histogram lies in the curving continuous lines that define the boundaries between topic currents. However, we do not have continuous data and so 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 honesty. If the user zooms in farther than the data resolution supports, our 'truth' as 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 topic weights. In this way, we can maintain a suitable level of truth 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.

5. USABILITY EVALUATION

To assess the potential value of our ideas in this prototype, we carried out a simple formative usability evaluation with two users. For the sake of comparison, we represented the same data both in ThemeRiver and in a histogram that we created using a spreadsheet. (See Figure 4.)

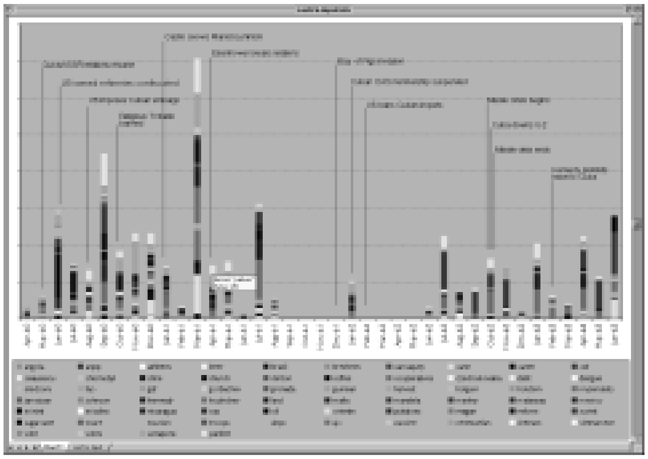

Figure 4: Histogram showing theme strengths in the Castro collection, April 1960-June 1963.

The data were the Castro collection described above. The histogram depicted thematic content by months, using the same values that drive ThemeRiver. 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 ThemeRiver. We then asked the participants questions about what they observed in the display. We captured verbal protocol during this discussion. We also asked participants to complete a short questionnaire.

Participants found ThemeRiver easy to understand, giving it an average rating of 2.5 on a scale of –3 to +3. They also found ThemeRiver to be useful, particularly for identifying macro trends, and gave it a rating of 2 on the same scale. They told us that it was less useful for identifying minor trends, because the curves tend to de-emphasize very small values. Users found that the connectedness of the river helped them follow a trend more easily over time than in the histogram. Asked whether they would be more likely to use the ThemeRiver or the histogram, both users answered that they were more likely to use ThemeRiver — one user more so than the other.

Users liked some features of the histogram and recommended adding them to ThemeRiver. One 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 in fact require that we interpolate between data points to produce the curves.

Although participants liked the abstraction to the whole collection and thus away from individual documents, both participants 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 to the text of each document on demand. They also wanted to select a current and see the documents that contributed to it.

6. FUTURE WORK

The current version of ThemeRiver is a demonstration prototype, developed to test the concept. We need to develop ways to build the event timeline automatically, instead of entering the data by hand as we did for the prototype.

From formative usability evaluation, we learned that users want to know more about the context of the river, especially how the content reflected in ThemeRiver compares to the full collection content. They also stressed the need for access to the documents that contribute to ThemeRiver 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.

7. ACKNOWLEDGMENTS

We gratefully acknowledge the contributions of our colleagues at Battelle to the development and testing of ThemeRiver. Special thanks go to Shelly Harris, Vern Crow, Feng Gao, Paul Whitney, Kris Cook, Dennis McQuerry, Jim Thomas, and Lou Martucci. This research was funded by Battelle Memorial Institute’s Information Synthesis Initiative.

8. REFERENCES

[1] Edwards, K.W. and Mynatt, E.D. Timewarp: Techniques for Autonomous Collaboration. In *Proceedings of CHI’97*, 218-225.

[2] Karam, G.M. Visualization Using Timelines. In *Proceedings of the 1994 International Symposium on Software Testing and Analysis*, 125-137.

[3] Kullberg, R.L. Dynamic Timelines: Visualizing the History of Photography. In *Proceedings of CHI ’96*, 386-397.

[4] Mackay, W. and Beaudouin-Lafon, M. Diva: Exploratory Data Analysis with Multimedia Streams. In *Proceedings of CHI’98*, 416-423.

[5] Nowell, L.T.; France, R.K.; Hix, D.; Heath, L.S.; and Fox, E.A. Visualizing Search Results: Some Alternatives to Query-Document Similarity. In *Proceedings of SIGIR ’96*, Aug. 1996, Zurich: ACM Press, 67-75.

[6] Plaisant, C.; Heller, D; Li, J; Shneiderman, B; Mushinlin, Rl and Karat, J. Visualizing Medical Records with LifeLines. In *CHI ’98 Summary*, 28-29.

[7] Tufte, E. R. *Visual Explanations: Images and Quantities, Evidence and Narrative*. Cheshire, CT: Graphics Press, 1997, 90-91.

[8] Wise, J.A.; Thomas, J.J.; Pennock, K; Lantrip, D; Pottier, M. Schur, A; and Crow, V. Visualizing the Non-Visual: Spatial analysis and interaction with information from text documents. In Card, S.K.; Mackinlay, J. D.; and Shneiderman, B. (ed.), *Readings in Information Visualization: Using Vision to Think*, San Francisco: Morgan Kaufmann, 1999, 442-450.