Analytic Worksheets: A Framework to Support Human Analysis of Large Streaming Data Volumes

Grace Crowder[2], Sterling Foster[2], Daniel M. Russell[1], Malcolm Slaney[1], and Lisa Yanguas[1]

[1] IBM Almaden Research Center, 650 Harry Rd., San Jose, CA, 95120
[2] U.S. Department of Defense, 9800 Savage Rd., Fort Meade, MD 20755
{gacrowder, ssfoster}@afterlife.ncsc.mil, daniel2@us.ibm.com, malcolm@ieee.org, lryangu@us.ibm.com

Abstract. Worksheets are a new user-interface framework to support analysis of streaming data by combining streaming data queries with visualization objects in a composable document framework. A worksheet lets users work at human speeds with large quantities of streaming data by creating a persistent, literate, dynamic document that flows data into analysis patterns of filters and visual presentations. The worksheet provides basic support for analysis created, as well as buffering and managing streaming data as it continually arrives.

1 Motivation and Approach

A significant problem most analysts face is the volume, velocity and variety of data they must manage and manipulate in a timely manner. Such pressure constantly works against careful and through analysis, creating a tension between completeness and publication timeliness for a comprehensive story based on the synthesis of this data and information.

Traditional intelligence analysis schemes have relied on stored data or information in databases, which analysts and other users subsequently access to make queries. Yet current analysts find themselves being buried under a constant onslaught of information pushes. Thus we focus on a way to work with the streaming aspect of data, a situation more closely resembling actual information flows in practice, where data constantly comes into the system, and where decisions are made on data while it is in motion rather than at rest. This adds to the complexity of the analytic task, but provides real benefits for more up-to-date and valid interpretations of the information stream.

In addition, collaboration within and among multiple users is an age-old problem. While there are many impediments to collaboration, an obvious one is the dearth of tools that truly support a collaboration environment as it would best serve users and their tasks. Analysts typically create a "sandbox" area while they are assessing data and information and creating a strategy to understand that data. Such models are often tentative and exploratory. They are typically not shared early on, but results are developed to a more finished form before being shared with others. However, once

M.F. Costabile and F. Paternò (Eds.): INTERACT 2005, LNCS 3585, pp. 950 – 953, 2005.

users have formed an assessment, it would be helpful to be able to show what specific pieces or aspects of data, analyses, thought processes, queries, etc. went into the user's making that particular assessment.

Our goals in designing analytic worksheets are to (a) provide tools to manage streaming data tasks, (b) to create and capture analysis patterns as a way to express the intent and context of a particular kind of analysis in progress, (c) form the basis for collaboration between analysts working on shared or similar problems.

A worksheet is a view of relevant data streams, packaging together data with contextualizing information that helps organize and communicate the work being done. It not only supports analysts in sharing information among themselves but also allows information to be shared with others involved in the analytic process. An analyst working on a particular problem can see in the worksheet what others working on similar or related issues have done, what queries they have levied on the system, what actions have been taken and how they manage incoming information (i.e., data streams).

2 Worksheets: A Way to Handle Large Volumes of Data for Analysis

Figure 1 shows an example of a worksheet. A worksheet is an easily authored, persistent, live document that is built up out of *inquiries* (persistent queries over streams of live data, continually filtering the stream for records that match their query specifications), *annotations* (user-editable text and graphics to document or comment on the analysis process or context), *visualizations* (linked to the data streams coming out of inquiries to filter and visually present the output of the inquiries), and *notifications* (that cause an email or IM to be sent out when a specified, exceptional condition in the data stream is reached). A user creates a hierarchy of these components by creating inquiries on specific data source streams, then attaching visualizations and notifications to that inquiry. Like any hierarchy, subsections of the worksheet can be minimized to hide details of the document that the reader does not want to see at the moment. In this way a worksheet seems very much like an outline tool, where inquiries form section headings with visualizations and notifications below as parts of the section.

One inquiry can produce a wealth of data for which a number of visualizations might be appropriate. A user creates an inquiry, embedding it in the worksheet, then specifies as many visualizations as necessary to understand the data, providing multiple views of the data that might filter or reorganize the collected data. In our prototype these visualizations include tabular views of records, geospatial maps, image viewers and simple graphical representations (e.g., line graphs and histograms); data sources include news wire stories, stock prices and realtime sensor data.

Intrinsic to the worksheet model is its ability to handle streaming data. The worksheet manages streaming data to provide a match between the incessant push of data and the need for users to slow down, pause and even back up in the collected data. That is, in order to operate with multiple, parallel inquiry streams, the worksheet buffers content streaming in from active inquiries to provide flow management and

received data storage. When working with live streaming data, the capability to pause the inquiry stream, extract a portion of the data, and back up to do a historical review of earlier data is critical.

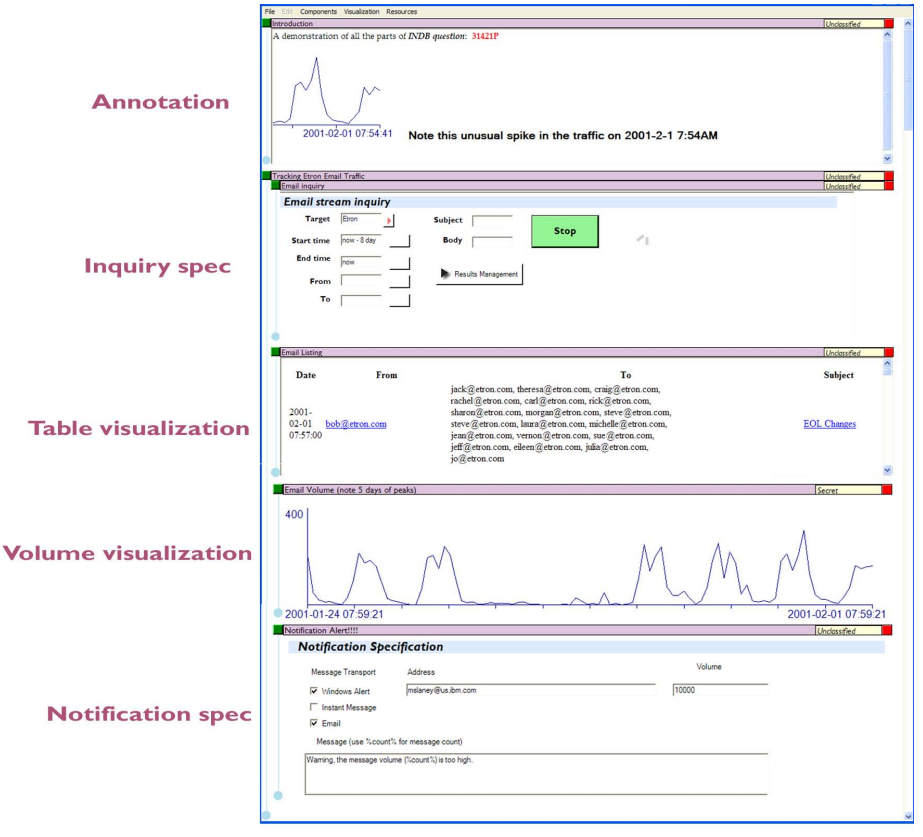

Fig. 1. This worksheet fragment shows an inquiry and two visualizations within the scope of the inquiry. Each visualization is linked to the inquiry, which in turn is updated by a stream of results coming from the Worksheet server. Worksheets provide basic capabilities to handle multiple live data streams with pause, flow management and the ability to back up for historical review of streamed data.

The worksheet promotes collaboration by allowing individual and groups of widgets to be selected, copied and pasted into a new worksheet. A portion of the worksheet, containing any number of text, inquiry, or visualization components can be copied as a group and sent to another user. A portion of the worksheet might address a particularly difficult question, and can be shared by simply copying and pasting into a new worksheet. Since the worksheet model is persistent, the entire structure of annotations, visualizations, inquiries and notifications can be copied and shared as a document.

3 Summary

Our worksheet model combines a few existing interface techniques: the persistent notebook, data flow graphs and component-style visualizations embedded in an organizing framework a la OpenDoc [6]. The Virtual Notebook [3] is a composite document holding interactive components in a persistent document-like object that can be edited, read, manipulated and monitored by the user.

The notebook was popularized in Mathematica [2], which provided a notebook interface as its primary interface model. Users combine text and graphics with mathematical expressions to display interactive results. If you change the definition of an equation, the display updates immediately showing its new values.

The worksheet-style interface has several advantages. It is *literate*, capturing the steps of the analysis, exposing intermediate work and assumptions. In this way the worksheet is a document that is meant to be read, like a book, instead of executed like a computer program [4, 5]. A worksheet is *dynamic*, continuously reflecting changes in the streaming environment as new information arrives.

The worksheet becomes a tool for sensemaking [1] when it captures the knowledge patterns of analysts. That is, a user can easily construct a tree of inquiry objects that import data from streams and then provide visualization tools to look at, examine and work with the data stream. The constructed worksheet is then a representation of the analytic framework, illustrating what works for this kind of problem.

Finally, worksheets are a way to do real analysis work over data streams; not simply collecting and organizing evidence, but also providing support for comparing evidence – pro and con. As data streams through the worksheets, we want to be able to create the best possible interpretation based on *currently* available information. The inherently live streaming nature of the worksheet approach allows users to keep these analyses up-to-date and accurate.

References

1. Russell, Daniel M., Stefik, Mark J., Pirolli, Peter, Card, Stuart. K. The cost structure of sensemaking. In Proc. of ACM INTERCHI'93 Conference on Human Factors in Computing Systems, 269—276 (1993)
2. Wolfram, Stephen *Mathematica: A System for Doing Math by Computer,* Addison Wesley (1990)
3. Burger, Andrew M., Meyer, Barry D., Jung, Cindy P., Long, Kevin B. The Virtual Notebook System, Proceedings of ACM Hypertext Conference, 395—402 (1991)
4. Slaney, Malcolm Interactive Signal Processing Documents, in *Symbolic and Knowledge-Based Signal Processing*. A. Oppenheim, H. Nawab (eds.), Prentice Hall, NJ (1992)
5. Knuth, Donald E. *Literate Programming*, Center for the Study of Language and Information Press, Stanford, CA (1992)
6. Feiler, Jesse; Meadow, Anthony *EssentialOpenDoc.* Addison Wesley, Reading, MA. (1996)