An Advanced Analytical Environment for Scientific Discovery within Continuous, Time-Varying Data-Streams

Andrew J. Cowell, Antonio P. Sanfilippo,
Susan L. Havre, Judi R. Thomson, Mark A. Whiting.
Pacific Northwest National Laboratory
Richland, WA, 99352, USA.
andrew@pnl.gov
(509) 375 4548

ABSTRACT
This paper discusses our recent work across a number of disciplines, leading to a concept for a next generation analytical environment for scientific discovery within continuous, time-varying data-streams. First, we have created a stream-processing engine that processes multiple streams of interest. An analyst, via a client interface, reviews the data-stream format and specifies upstream filtering to define stream tokens of interest, leading to a highly specialized collection of time-variant material. We envision using this collection to drive an existing system that visualizes thematic variations over time across a corpus of information. This 'ThemeRiver™' helps analysts discern trends, relationships, anomalies, and structure in the data. Further, we make use of a number of technologies that allow us to investigate these elements in ambient environments that surround the user, placing them within their data. We discuss the HI-SPACE (Human Information Space) as a tool for bringing together the most desirable aspects of both physical and electronic information spaces to enhance the ability to interact with information, promote group dialog, and to facilitate group interaction with information to solve complex tasks. Here, we introduce a concept that combines these approaches to produce an advanced analytical environment for data stream analysis that provides a collaborative, ambient environment for scientific discovery in data-streams.

Author Keywords
Information Visualization, Data Streams, Human Computer Interaction, Novel Interaction Paradigms.

ACM Classification Keywords
H.5 Information Interfaces and Representation (HCI)

INTRODUCTION
The term 'data-stream' is an increasingly overloaded expression. For our purposes we define data-stream as a series of data (e.g. credit card transactions arriving at a clearing office, cellular phone traffic or environmental data from satellites) arriving in real time. We call each data element in the stream a token. The complexity of these tokens ranges from simple (e.g. characters in a sentence: "T H I S I S A S T R E A M…") to extremely complex (e.g. a detailed transaction record). The volume of data-streams is usually massive, and while each individual token may be rather uninformative, taken as a whole they describe the nature of the changing phenomena over time.

Data-streams differ from conventional stored relations in many ways. They have no width or flow boundaries, meaning that there is no control over the total amount of data flowing, or differences in flow volume arriving at any particular moment. They are time-varying and unpredictable. Flow can start or stop at any point and the number of tokens per unit time that are delivered to a receiver vary. We also have no control over the order in which data items arrive. Some data-streams provide tokens in order, while others do not. Finally, because of the volume of data after processing tokens from the data stream, the tokens typically are discarded and cannot (usually) be retrieved for further analysis (i.e. they are non-persistent).

Data-streams are essential parts of the scientific discipline, intrinsic to a growing number of research areas. 'Knowledge workers' who find themselves working in these areas are in dire need of tools that help them deal with the substantial volume and high complexity that modern streamed data possesses.

Our goal in this paper is to put forth a concept for a future generation analytical environment for scientific discovery within data-streams. We take a component approach, utilizing prototypical elements designed under different auspices for diverse purposes and describe an approach to bring them together to provide an ambient environment for collaborative study and a platform for data-stream research.

STREAM PROCESSING COMPONENT

The overall concept of our stream-processing engine (shown in Figure 1) is to provide the ability to drastically reduce incoming data-streams to more manageable levels by allowing the knowledge worker to implicitly define filters[1] that restrict content to only those tokens of intellectual value. In our prototype (built around the domain of detecting significant change in streaming data), this reduced stream is then routed through a set of algorithms that produce signatures (a compressed representation of the original token).

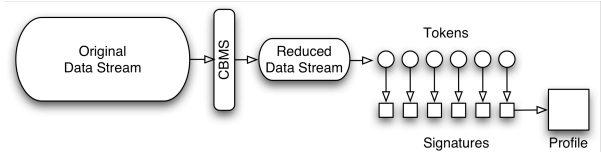

Figure 1: Data-Stream Processing Engine

A signature expresses the semantic content of the data sub-stream it encodes with reference to topics that are discovered through an unsupervised classification model [2] Such a classification model is augmented with a process of ontological annotation that identifies relevant entities and relations among them in terms of reference generic and domain specific ontologies. The topics, entities and relations discovered are then utilized to provide users with an information rich visualization of the data stream.

As signatures are generated, they are consumed into a descriptive profile (a representation of the *status quo* of the reduced stream). On a token-by-token basis, the profile may grow or remain the same depending on what that particular signature adds to the current knowledge of this stream. After a user-defined training period, new signatures from arriving tokens in the stream are compared against the profile and evaluated for novel content that the knowledge worker may be interested in.

For the purposes of our overarching concept here, we rely on the architecture to reduce the initial data-stream to a manageable and interesting set, providing a dynamic corpus of documents for the visualization component.

VISUALIZATION COMPONENT

One goal in exploratory information visualization is to present information so that users can easily discern patterns. These may reveal trends, relationships, anomalies, and structure in the data, and could potentially help confirm knowledge or hypotheses. Patterns may also raise unexpected questions leading users to new insights. The challenge is to create visualizations that enable users to find patterns quickly and easily. Researchers at the Pacific Northwest National Laboratory have developed a suite of visualization tools that help the knowledge worker in these tasks [3]. Here, we present ThemeRiver™[2] [4] (Figure 2). 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. The main task at this stage is to refine the ingest mechanism to allow for dynamic data stores (ThemeRiver currently works on a static set of documents). Other visualization tools provide insights on the content makeup of each theme and capture significant links across themes, e.g. in terms of the entities and relations discovered through ontological annotation.

By associating a mechanism for collecting and simplifying data-streams with a themed visualization tool, we shall be taking a step towards providing knowledge workers with the types of tools they require for successful stream processing. While this is certainly a step in the right direction, we are intrigued by the notion of being subsumed by the data, being perceptionally linked to its flow. For these reasons we consider a third component, a means to engulf the user in the data.

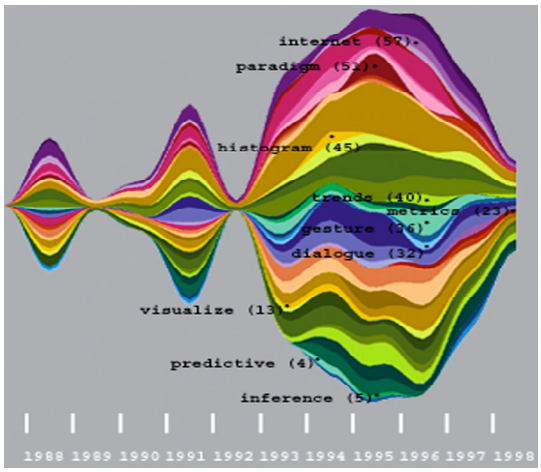

Figure 2: ThemeRiver

AMBIENT COMPONENT

The key to developing the next generation human to information interface is to move beyond the limitations of small computer monitors as our only view into the electronic information space and keyboards and mice as the only interaction devices. Our physical information space, which includes walls, tables, and other surfaces, should now be our view into the electronic information space. People perform physical interactions with information every day by picking up a book, building a model, or writing notes on a page. Similar interactions need to be developed for electronic information. Providing these types

[1] [1] discusses more details of the filtering component.

[2] ThemeRiver is a registered trademark of the Pacific Northwest National Laboratory.

of interactions in the electronic world would allow us to interact more quickly, naturally, and hopefully more effectively in the broader context of information exploration. For these reasons, Battelle, in association with the HITLab, have created the Human Information Space (HI-SPACE) [5] (Figure 3). This system utilizes knowledge from many areas of research, including Psychology, Human-Computer Interaction, Virtual Reality, Kinesiology, and Computer Science, to create a physical workspace that blurs the boundaries between physical and electronic information. The most desirable aspects of both the physical and electronic information spaces are used to enhance the ability to interact with information, promote group dialog, and to facilitate group interaction with information to solve complex tasks.

Figure 3: The Human-Information Space (HI-SPACE)

CONSOLDIDATED SYSTEM

The consolidated environment takes each of these elements together to produce a rich environment for collaborative investigation. Knowledge workers standing around the HI-SPACE could observe the themes of the selected streams flow, changing dimensions as new token influence the *status quo*. Monitors covering the walls could hold representations of individual data items or other miscellaneous statistics. They could also provide expressions of other researchers, connected in from a remote location. We envision the HI-SPACE itself holding the ThemeRiver image, but enhanced to allow users to use hand gestures to indicate topics of interest. They may use phicons [6], physical objects registered with the HI-SPACE as placemarkers. These could be placed as stopblocks on themes that are not considered essential to the current study, or in some alternative manner so to affect the themes presented. Usability studies will be required to ascertain the usefulness of an ambient approach to such scientific discovery, but we believe that by becoming more intrinsically involved with the data, scientists may be able to uncover unique findings, previously undiscovered. An artist's rendition of our concept can be seen in Figure 4.

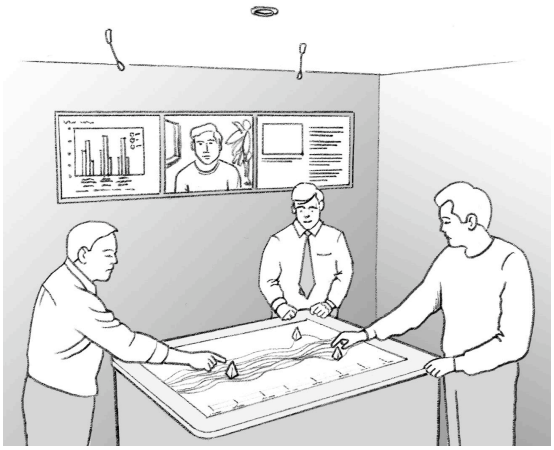

Figure 4: Next Generation Environment for Scientific Discovery in Data-Streams (Concept)

SUMMARY

We have presented our vision for a next generation analytical environment for scientific discovery within data-streams. We intend to build this prototype via a component approach, utilizing elements of our research portfolio in Information Analytics, Rich Interaction Environments and Knowledge Engineering.

ACKNOWLEDGMENTS

The authors would like to thank their colleagues across a number of departments at PNNL that made these individuals components possible.

REFERENCES

1. I.Gorton, Justin Almquist, Nick Cramer, Jereme Haack, Mark Hoza. 2003. "An Efficient, Scalable Content-Based Messaging System", in Proc. 7th IEEE International Enterprise Distributed Object Computing Conference (EDOC 2003), pg. 278-285, Brisbane, Australia, September, 2003.
2. Wise, J. A., J. J. Thomas, *et al*. 1995. Visualizing the Non-Visual: Spatial Analysis & Interaction with Information from Text Documents. IEEE Information Visualization. IEEE Press, Los Alamitos, CA.
3. Pacific Northwest National Laboratory's IT Showcase http://showcase.pnl.gov/show?it/triver-prod webpage.
4. Havre SL, Hetzler BG, Whitney PD and Nowell LT. 2002. "ThemeRiver: Visualizing Thematic Changes in Document Collections". IEEE Transactions on Visualization and Computer Graphics. 8(1):9-20.
5. RA May, JJ Thomas, RR Lewis, SD Decker. 1998. "Physical Human Information Interface Workspace", Proceedings of Western Computer Graphics Symposium '98, April 23-26, pp 25-31.
6. Ullmer, B. and Ishii, H. 1997. "The metaDESK: Models and Prototypes for Tangible User Interfaces." UIST '97 Proceedings, pp. 209-10.