Visual Data Mining

Pak Chung Wong
Pacific Northwest National Laboratory

Seeing is knowing, though merely seeing is not enough. When you understand what you see, seeing becomes believing. A while ago scientists discovered that seeing and understanding together enable humans to glean knowledge and deeper insight from large amounts of data. The approach integrates the human mind's exploration abilities with the enormous processing power of computers to form a powerful knowledge discovery environment that capitalizes on the best of both worlds. The technology builds on visual and analytical processes developed in various disciplines including scientific visualization, data mining, statistics, and machine learning with custom extensions that handle very large, multidimensional, multivariate data sets. The methodology is based on both functionality that characterizes structures and displays data and human capabilities that perceive patterns, exceptions, trends, and relationships. Here I'll define the vision, present the state of the art, and discuss the future of a young discipline called visual data mining.

The vision

The vision of a visual data mining system stems from the following principles: simplicity, user autonomy, reliability, reusability, availability, and security.

A visual data mining system must be syntactically simple to be useful. Simple doesn't mean trivial or nonpowerful. Simple to learn means use of intuitive and friendly input mechanisms as well as instinctive and easy-to-interpret output knowledge. Simple to apply means an effective discourse between humans and information. Simple to retrieve or recall means a customized data structure to facilitate fast and reliable searches. Simple to execute means a minimum number of steps needed to achieve the results. In short, simple means the smallest, functionally sufficient system possible.

A genuine visual data mining system must not impose knowledge on its users, but instead guide them through the mining process to draw conclusions. Humans should study the visual abstractions and gain insight instead of accepting an automated decision.

A reliable visual data mining system must provide estimated error or accuracy of the projected information for each step of the mining process. This error information can compensate for the deficiency that imprecise analysis of data visualization can cause.

A reusable visual data mining system must be adaptable to a variety of systems and environments to reduce the customization effort, provide assured performance, and improve system portability.

A practical visual data mining system must be generally and widely available. The quest for new knowledge or deeper insights of existing knowledge cannot be planned. It may mean a portable system through telelinks or an embedded (local) system within the information domain. This requires that the knowledge received from one domain adapt to another domain through physical means or electronic connections.

Finally, a complete visual data mining system must include security measures to protect the data, the newly discovered knowledge, and the user's identity because of various social issues.

So far I've ignored discussing the underlying visualization and mathematical techniques of visual data mining. This is partly because of the space limit and partly because of the steady and incremental technological advancements in the field of visual data mining. You can find samples of the latest technologies in this special issue. Although no one involved in this exciting field has all the technical solutions today, everyone is fully aware of the grand challenges ahead.

Current state of the art

Visualization has been used routinely in data mining as a presentation tool to generate initial views, navigate data with complicated structures, and convey the results of an analysis. Generally, the analytical methods themselves don't involve visualization. The loosely coupled relationships between visualization and analytical data mining techniques represent the majority of today's state of the art in visual data mining. The process sandwich strategy, which interlaces analytical

0272-1716/99/$10.00 © 1999 IEEE

processes with graphic visualization, penalizes both procedures with each other's deficiencies and limitations. For example, because an analytical process can't analyze multimedia data, we have to give up the strengths of visualization to study movies and music in a visual data mining environment.

Perhaps a stronger visual data mining strategy lies in tightly coupling the visualizations and analytical processes into one data mining tool. Letting human visualization participate in an analytical process' decision-making remains a major challenge. Certain mathematical steps within an analytical procedure may be substituted by human decisions based on visualization to allow the same analytical procedure to analyze a broader scope of information. Visualization supports humans in dealing with decisions that can no longer be automated. This results in a tightly coupled visual data mining environment that truly takes advantage of the strengths of all worlds.

The future

All signs indicate that the field of visual data mining will continue to grow at an even faster pace in the future. In universities and research labs, visual data mining will play a major role in physical and information sciences in the study of even larger and more complex scientific data sets. It will also play an active role in nontechnical disciplines to establish knowledge domains to search for answers and truths. For example, there may exist standard "man" pages for our favorite visual data mining functions on our Unix system. An advanced form of scatterplot matrix may substitute for the use of covariance and regression in statistics studies. National standards will be developed to govern the functionality and resources of visual data mining.

In industries and households across the country, visual data mining will be embedded in public utilities and home appliances. Many searching references—such as the yellow pages, dictionaries, and even newspapers—will have visual mining capability. There may be computer chips dedicated to support visual data mining activities. The term visual data mining will be included in school textbooks and literature. Audio- or haptic-based substitutes will help the visually impaired. Our imagination is the only limit of the future.

About the articles

This special issue on visual data mining attracted high-quality submissions from England, Germany, and the United States. All of the articles presented interesting and promising research in visual data mining. Unfortunately, there's only room to include a few articles.

Hinneburg, Keim, and Wawryniuk introduce a novel clustering algorithm on large amounts of high-dimensional data. Visualization techniques instead of automated decisions guide the recursive partitioning of the new clustering algorithm. This is a major step towards the goal of a tightly coupled visual data mining environment.

Ribarsky et al. present a clustering algorithm for very large data sets. The new technique enables clustering of large amounts of data with a fast interactive response time. This provides continuous interactions between man and machine during the data unfolding process. Because of the layered design of the visualization and clustering processes, it's considered a loosely coupled visual mining system.

Rohrer, Ebert, and Sibert describe a shape-based visualization system to support data mining of text. The text information is mapped to document vectors before it's visualized using implicit surface modeling techniques. The system supports querying articles of a corpus by matching the vector's shape through visualization. Zoom views support data drilling of the document text.

Conclusion

This issue showcases an exciting field where people turn seeing into knowing, believing, and eventually human insights. I believe the vision defined here can be reached and the proposed tasks accomplished. As the articles in this issue show, both the loosely and tightly coupled visual data mining systems perform well in certain domains and environments. Scientists and engineers will continue to explore new ground and find new applications in this young discipline. As for the future, I see the advancement of visual data mining resembling the rapid growth of personal computers in our society. The active participation of humans and the decisions based on visualization combine the art of human intuition and the science of mathematical deduction, forever changing the landscape of data analysis. ■

Acknowledgments

The Pacific Northwest National Laboratory is operated for the US Department of Energy by Battelle Memorial Institute under contract DE-AC06-76RLO 1830.

Pak Chung Wong *is a senior research scientist in the Synthesis, Analysis, and Visualization of Information group at the Pacific Northwest National Laboratory in Richland, Washington, where he performs research and development on scientific computation and information technology. His research interests include visualization, data mining, scientific data abstraction, steganography, and wavelets. He received a PhD in computer science from the University of New Hampshire.*

Readers may contact Wong at the Pacific Northwest National Laboratory, 902 Battelle Blvd., P.O. Box 999, MSIN: K7-28, Richland, WA 99352, e-mail pak.wong@pnl.gov.