Microsoft Research

Microsoft External Research

Making Sense of Data Overload: An Innovative Approach to Progressive Data Analysis

With the help of Microsoft Research, USC's Cyrus Shahabi is working to help scientists manage data to provide better results, reliability and overall accuracy of data analysis.

Improvements in data-gathering technologies have given companies and the scientific community the ability to collect massive amounts of information. Although having that much data opens the doors to understanding our world with greater depth, historically the sheer size of these datasets has made it nearly impossible for scientists and academic researchers to make sense of it all without reducing the size of the data first.

The problem is, data reduction results in data loss, making it necessary to use modeling, interpolation or simulations to reconstruct the results of the information that has been gathered — which can be time consuming and less accurate than dealing with a full dataset.

Scientists have struggled with competing interests, speed versus accuracy versus costs. How do you make it better? That's the question Microsoft Research posed. Designed to engage software designers, academics and scientists, Microsoft Research encouraged ingenuity and creativity not only in tackling ways to make technical work more manageable, but also in making the process more user friendly.

"At Microsoft Corp., we're always looking for ways to solve these problems. That leads us to engage with members of the scientific and academic communities regularly," said Dan Fay, program manager of the Microsoft External Research Group — the arm of Microsoft Research that works most closely with the academic and scientific research communities. "What we hear during these conversations is that there are problems finding adequate software to manage the growing tonnage of data they have."

Recognizing the inherent limitations of most data analysis software, Professor Cyrus Shahabi of the University of Southern California (USC) began to think of new ways to analyze large amounts of information. "The entire idea of e-science is to find ways to get through data effectively and quickly. Unfortunately, with most of the current programs, there is too much data for scientists or analysts to do productive analysis," Shahabi said. Then it came to him: In signal processing they are able to manipulate large quantities of data without the problematic issues involved in other forms of data analysis.

"In signal processing, they use different types of compression tools to reduce the size of the large datasets to save space and/or communication cost. This is achieved by compactly storing the main patterns in the data in such a way that the dataset can be reconstructed back in its entirety from those patterns, with minimal loss of accuracy," he said. However, in data analysis the focus is on reconstructing only a portion of data that is under investigation. Hence, the compression techniques cannot be applied directly to the data analysis problem. In fact, as Shahabi said, "You compress for speed, not space. The problem is speed. Storage is cheap."

So, understanding that, Shahabi began to develop methods for scientific data analysis based on the benefits he found in signal processing. The idea was to "transform" the data using a signal processing tool (i.e., wavelet transformation) but not "compress" it. Instead, at the query time, when the system knows which portion of the dataset is needed for the analysis, that portion is reconstructed on the fly. In addition, the recon-

Fast Facts

Project: ProDa for Scientific Data Analysis

Project Principal: Professor Cyrus Shahabi

Partner: University of Southern California

Web Site:
http://infolab.usc.edu/projects/proda

Profile:
When Professor Cyrus Shahabi of the University of Southern California decided to tackle the problem of complex data analysis, he was confronted by the limitations of current software. Realizing what an impediment this was for businesses and the scientific community, he began to explore alternative forms of analysis. When he came across signal processing and wavelet compression, he knew he was onto something, and ProDA was born. Since creating ProDA, NASA's JPL and Chevron have had major successes using the program to manage their huge datasets. With the help of Microsoft Research's Smart Client initiative, Shahabi was able to bring ProDA to the next level by making it more compatible with XML, Microsoft Excel, text files, and many more formats. All these changes have made ProDA more accessible and user friendly.

Microsoft Research Initiative

As sciences become more data intensive, computational technologies are beginning to transform scientific research. Tools for data gathering, mining, analysis and visualization are becoming integral to the practice of science, often yielding dramatic productivity gains. To help create e-science solutions, Microsoft Research encouraged scientists to incorporate advanced technologies within their research. This year, Shahabi took on that challenge with his innovative software, ProDA. By making his software more compatible with applications such as Microsoft Excel, Shahabi enhanced his program's usability, allowing broader business and scientific access.

ProDA received support from the Microsoft External Research Group within Microsoft Research, which focuses on advancing e-science by collaborating with the world's foremost researchers in academia, industry and government to move research in new directions across nearly every field of computer science, engineering and general science.

e-Science Initiative:
http://www.microsoft.com/science
Microsoft Research:
http://research.microsoft.com

Microsoft

struction occurs at the resolution level required by the user. For example, if the user is interested in monthly temperature across the Earth, there is no reason to reconstruct the temperature data all the way to their original hourly values. Shahabi refers to this method as "query compression" as opposed to "data compression," which he identifies as his most fundamental research contribution. The wavelet transformation of queries not only results in faster response time but also enables new query answering features such as providing quick approximate answers and progressive answers (i.e., the query result gets more and more accurate over time).

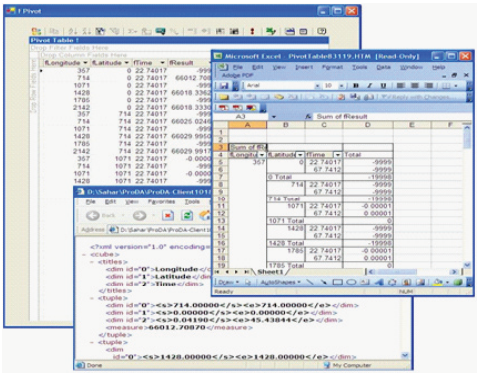
In 1999 Shahabi developed the fundamental ideas behind his wavelet-based query compression method, which later led to the development of Progressive Data Analysis (ProDA). Right from the start ProDA gained the interest of and became an asset to NASA's Jet Propulsion Laboratory (JPL), helping it analyze massive amounts of temperature data from all over the planet. Meanwhile, the theoretical foundation of applying wavelets to query processing attracted funding from the National Science Foundation, and eventually a second round of funding from NASA to refine the ProDA's processing efficiency gained the attention of Chevron Corp. "Chevron uses multiple high-rate sensors that generate terabytes worth of data, so ProDA's automatic storage and analysis methods helped them understand what was happening in their oil reservoirs," Shahabi said. Chevron is planning to use ProDA to sort through mountains of oil production and water injection data to increase oil production at some of its fields.

Perfect Timing

During his work with Chevron and JPL, Shahabi noticed that the work everyone was doing, and in particular the systems they were using, were all powered by Microsoft Office applications. He realized that with the help from Microsoft Research he'd be in a position to bring ProDA to the next level — making it more compatible with Microsoft Office Excel.

By taking steps to make ProDA more accessible, user friendly and compatible with Excel, the program has been better suited to analyze real-world data — making it even more useful than before. As the usability of the program has advanced, it has simultaneously opened up to broader applicability for industries and companies beyond science.

"Before we partnered with Microsoft Research, honestly, we had a hard time selling the benefits of ProDA to potential users. But now, we are able to position it on a platform that everyone is familiar with and uses, Microsoft Excel," Shahabi said. In particular, ProDA takes advantage of the PivotTable component of Microsoft Excel, which
allows a good deal of flexibility when performing on-the-fly analysis of large sets of data. The PivotTable can be used to work with both basic and complicated calculation modules independent of the spreadsheet's layout. Shahabi said his goal in further developing ProDA was to create a deeper sense of interactivity for the application and making it easy for companies to understand. Shahabi said, "Now the users can analyze the data, make comparisons, detect patterns and relationships, and discover trends. With its extensive set of export functionalities, ProDA can be connected to almost any application. At any time, users can export their query results to XML, Excel, text files, and many more formats."



Fay is impressed with ProDA's innovative approach: "By using the wavelet approach, you're able to see trends quickly but still access the full range of data. We see a lot of interesting proposals, but this was the perfect blend of innovation and real results."

One feature that sets ProDA apart from any other analysis software is its ability to compute complex statistical queries on the fly — it is the only data analysis software that is able to do so. By blending with Microsoft Excel, ProDA is able to appeal to a broad range of users; and recently, Shahabi added a deeper layer of functionality by making the software capable of advanced geospatial visualization when working with applicable datasets.

By creating a more visually based component, users are able to exchange results with others they are working with while literally viewing data in relation to its geographical space. The multilevel functionality that has been built-in with each evolution of ProDA has dramatically improved performance and user experience while also decreasing training costs. All of this could be a benefit to companies as well as members of the scientific and academic communities.

Next Steps

Shahabi says the support from Microsoft has helped make ProDA more accessible and user friendly, which, in turn, he hopes will appeal to more industries and others within the scientific and academic communities. Recently NASA's JPL approached Shahabi to use ProDA's enhanced data retrieval and visual functionality for a worldwide oceanographic temperature mapping. It hopes that with ProDA it will be able to better predict major oceanic events such as hurricanes and typhoons. To date there are no plans to make ProDA commercially available; however, Shahabi says it isn't something he's ruling out.

© 2008 Microsoft Corporation. All rights reserved. This case study is for informational purposes only. MICROSOFT MAKES NO WARRANTIES, EXPRESS OR IMPLIED, IN THIS SUMMARY. Microsoft is a registered trademark of Microsoft Corporation in the United States and/or other countries. The names of actual companies and products mentioned herein may be the trademarks of their respective owners.