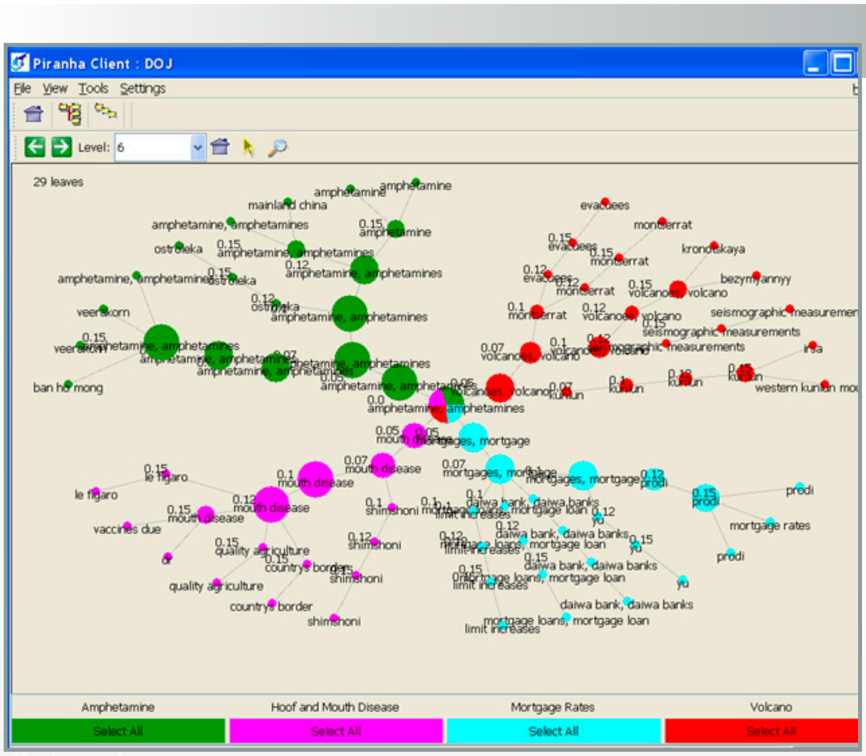
Agent-Based Software for Gathering and Summarizing Textual and Internet Information

UT-B IDs 200101031, 200401368, 200601759, 200902235, 200902377



Technology Summary

ORNL's Piranha solves the challenge most users face: finding a way to sift through large amounts of data that provide accurate and relevant information. This requires software that can quickly filter, relate, and show documents and relationships. Piranha is JavaScript search, analysis, storage, and retrieval software for uncertain, vague, or complex information retrieval from multiple sources such as the Internet. With Piranha, researchers have pioneered an agent approach to text analysis that uses a large number of agents distributed over very large computer clusters. Piranha is faster than conventional software and provides the capability to cluster massive amounts of textual information relatively quickly due to the scalability of the agent architecture.

While computers can analyze massive amounts of data, the sheer volume of data makes the most promising approaches impractical. Piranha allows advanced textual analysis to be accomplished with unprecedented accuracy on very large and dynamic data. For data already acquired, this design allows discovery of new opportunities or new areas of concern. Piranha has been vetted in the scientific community as well as in a number of real-world applications.

Piranha's Capabilities

- Finding Similar Documents: After selecting a document of interest, users can quickly find other similar documents.
- Sampling Documents: A set of documents usually contains common themes or topics. Representative documents from these themes can be found quickly and presented to an analyst.
- Classifying Documents: A set of representative documents can be used by an analyst to define a topic of interest, and then related documents can be added to that set.

Advantages

- More effective at collecting and summarizing large amounts of information from multiple sources
- Clustering technique compares and stores similar information and provides a visual display

Potential Applications

- Text mining
- Information "sense-making"
- Document organization
- Classification

Patents

Thomas E. Potok and Joel W. Reed, *Agent-Based Method for Distributed Clustering of Textual Information,* U.S. Patent 7,805,446, issued September 28, 2010.

Thomas E. Potok, Mark T. Elmore, Joel W. Reed, Nagiza F. Samatova and Jim N. Treadwell, *System for Gathering and Summarizing Internet Information,* U.S. Patent 7,072,883, issued July 4, 2006.

Thomas E. Potok, Mark T. Elmore, Joel W. Reed, Jim N. Treadwell, and Nagiza F. Samatova, *System for Gathering and Summarizing Internet Information,* U.S. Patent 7,315,858, issued January 1, 2008.

Y. Jiao and T. Potok, *Dynamic Dimensionality Reduction for Data Stream Analysis,* U.S. Patent Application 12/072,723, filed February 28, 2008.

B. Beckerman, R. Patton, and T. Potok, *Method for Learning Phrase Patterns from Textual Documents,* U.S. Patent Application 61/310,351, filed March 4, 2010.

R. Patton and T. Potok, *Detecting Temporal Precursor Words in Text Documents Using Wavelet Analysis,* U.S. Patent Application 61/331/941, filed May 6, 2010.

Lead Inventor

Thomas E. Potok
Computational Sciences and Engineering Division
Oak Ridge National Laboratory

Licensing Contact

David L. Sims
Technology Commercialization Manager, Building, Computational, and Transportation Sciences
UT-Battelle, LLC
Oak Ridge National Laboratory
Office Phone: 865. 241.3808
E-mail: simsdl@ornl.gov

Oak Ridge National Laboratory
MANAGED BY UT-BATTELLE FOR THE DEPARTMENT OF ENERGY
PARTNERSHIPS
ORNL 2010-G00618/jcn