Enchilada: Data Mining Software for Atmospheric Science

Jamie F. Olson, David R. Musicant, Leah E. Steinberg
Carleton College, Northfield, MN 55057

July 6, 2006

Introduction

Enchilada (Environmental Chemistry through Intelligent Atmospheric Data Analysis) is a software system in development that enables data mining on atmospheric data. Enchilada enables the analysis of atmospheric *mass spectra* and *time series* by providing a variety of data mining and visualization tools.

Data in Enchilada can come from a variety of sources. To date, the project has focused heavily on data from an Aerosol Time Of Flight Mass Spectrometer (ATOFMS) – a device used for gathering atmospheric data in a real-time manner. An ATOFMS obtains particles from ambient air, one at a time, and splits them apart one at a time with lasers. Different components of each particle take varying amounts of time to cross a fixed-distance chamber, depending on the mass of the component. By measuring the time it takes the fragments to make the journey, the approximate chemical composition of the orignal aerosol particle can be determined. This data is represented as a mass spectrum, which is a high dimensional histogram of the various elemental components that made up the particle. In its raw form, a typical ATOFMS dataset is high dimensional (60,000 dimensions), and consists of several million particles. After pre-processing and reducing dimensions to those which are scientifically meaningful, several hundred dimensions are still present.

Another source of data for Enchilada is time series data, such as wind direction or concentration of mercury in air over time. Such data can be analyzed in conjunction with mass spectra as described above.

Work on Enchilada is a joint effort between the University of Wisconsin-Madison and Carleton College, and the software represents an ongoing collaborative effort on the parts of computer scientists and atmospheric scientists at both institutions. The design of Enchilada is being heavily driven by the needs of the atmospheric scientists on the project: the goal is to deliver a software system allowing them to do innovative data mining without support from programmers. Furthermore, it is intended that Enchilada will allow such scientists to incorporate their valuable *domain knowledge* when carrying out their analyses.

Contributions

Enchilada is an open-source project, hosted on SourceForge. Unlike previous efforts at software systems for mass spectrum analysis that aim to provide programming libraries [2], Enchilada provides atmospheric scientists with an interactive graphical interface for access to data mining techniques. To date, Enchilada incorporates a wide variety of features,

including labeling, visualization, time-series aggregation, and clustering. Screen shots of Enchilada may be found at http://sourceforge.net/project/screenshots.php?group_id=133381, or alternatively by visiting http://sourceforge.net/projects/edam-enchilada/ and clicking on the "Screenshots" link.

Labeling [6] allows the user to be able to see what elements may be present in a particular spectrum. An aerosol particle typically consists of a combination of a variety of elements, and so it is of use to the practitioner to be able to observe what chemical elements compose it. Enchilada allows visualization of mass spectra and associated elemental labels.

Enchilada also facilitates clustering of aerosol particle data. Clustering algorithms we have implemented thus far include Art-2A (historically a popular clustering algorithm with chemists), k-means, k-medians, and BIRCH [8] (for datasets too large to fit in memory). Our distance metrics for clustering are comprised of Euclidean squared, Manhattan, and dot-product distances. In implementing the k-medians clustering technique on the ATOFMS data, we have developed a modification for normalization under the Manhattan distance metric. The Manhattan Normalization algorithm [3] that resulted from our efforts normalizes multi-dimensional data without skewing the relative locations of the data points in space. One of our near-future goals is to consult with the atmospheric scientists on the project to define a distance metric that better reflects the nature of the data that we handle, such as one which utilizes domain knowledge in weighting dimensions.

The interface that Enchilada provides places emphasis on presenting information efficiently and understandably to the researcher. This is of particular importance in the analysis of results achieved from clustering. Enabling atmospheric scientists to evaluate the quality and content of clusters on millions of points with hundreds of dimensions is a challenging task. We are therefore developing and implementing a histogram-based visualization technique to facilitate such efforts [1].

Enchilada also implements importation and visualization of time series data, and allows the user to take two time series of different resolutions and synchronize them for comparative analysis. The software also allows a collection of mass spectra to be converted to a collection of time series, one series for each dimension. This allows "apples to apples" comparisons between a particular dimension of ATOFMS data, for example, and time series data measured at the same time by other equipment (such as ambient mercury concentration). We are currently working to automate time series anomaly detection within Enchilada. This will allow users to discover "atmospheric events," such as a the appearance of an unexpected plume of substances in the air.

Technical Specifications

Enchilada is a Java application compliant with Java 2 SE 5.0, developed using Eclipse. Enchilada uses Microsoft SQL Server as its database.

Our organization of aerosol particles imported into Enchilada uses a somewhat unusual collection system. The particles are gathered into collections, and collections may have subcollections in a tree-like hierarchy. Unlike a file system, however, where each folder contains either more folders or files directly within that folder, every collection in Enchilada recursively and automatically contains all the particles contained in each of its subcollections. This design allows the user to partition collections into multiple sub-collections (such as when clustering), yet also seamlessly treat the original collection as a whole for future analysis. The software allows manipulation of collections such as copying, pasting, and combining. New collections may be created by a user, or on importation of new data, or via clustering or aggregation into time-series.

To support collection inheritance as well as other functions of Enchilada, the database is

designed with flexibility in mind. Enchilada incorporates a highly flexible database structure that can accommodate new datatypes defined by Enchilada users. This allows Enchilada to be easily expanded for use with any sort of mass spectrum or time series data. New datatypes can be specified using an XML importation system, resulting in the creation of new tables in the database on the fly. Common datatypes are built into the database along with tables necessary for organization. Data of the most-often used datatypes can be imported with easy dialog boxes, and data of any type can be imported using the XML importer.

In addition to providing data mining techniques for exploring the data, we have also implemented a rudimentary querying system within Enchilada to allow the atmospheric scientists to obtain a variety of summary statistics on their data.

Conclusion

Enchilada is a software system to facilitate the data mining and analysis of atmospheric data. It integrates traditional data mining algorithms with cutting edge research techniques to deal with the idiosyncrasies of atmospheric data, while providing it all in an interface that atmospheric scientists can use without the assistance of data miners or programmers.

Work on Enchilada is supported by NSF ITR grant IIS-0326328 and by Carleton College.

References

[1] *The Visual Display of Quantitative Information*. Graphics Press, 1992.

[2] Allen, J. O. *YAADA: Software Toolkit to Analyze Single-Particle Mass Spectral Data*, versions 1.3 amd 2.0 ed. Arizona State University, Tempe, AZ, October 2005.

[3] Anderson, B. J., Gross, D. S., Musicant, D. R., Ritz, A. M., Smith, T. G., and Steinberg, L. E. Adapting k-medians to generate normalized cluster centers. In *Proceedings of the Sixth SIAM International Conference on Data Mining*.

[4] Anderson, B. J., Musicant, D. R., Ritz, A. M., Ault, A., Gross, D. S., Yuen, M., and Gaelli, M. User-friendly clustering for atmospheric data analysis. Tech. rep., Carleton College Computer Science, 2005.

[5] Carpenter, G., Grossberg, S., and Rosen, D. Art 2-a: An adaptive resonance algorithm for rapid category learning and recognition. *Neural Networks* (1991).

[6] Huang, Z., Chen, L., Cai, J.-Y., Gross, D., Musicant, D. R., Ramakrishnan, R., Schauer, J. J., and Wright, S. J. Mass spectrum labeling: Theory and practice. In *Proceedings of the Fourth IEEE International Conference on Data Mining*.

[7] Ramakrishnan, R., Schauer, J. J., Chen, L., Huang, Z., Shafer, M., Gross, D. S., and Musicant, D. R. The edam project: Mining atmospheric datasets. *International Journal of Intelligent Systems* (2005).

[8] Zhang, T., Ramakrishnan, R., and Livny, M. BIRCH: an efficient data clustering method for very large databases. In *Proceedings of the 1996 ACM SIGMOD International Conference on Management of Data*.