

Computer-Chemie-Centrum
TORVS Research Team

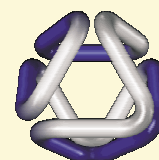
ChemVis

Interactive Datamining and Information Visualization

Frank Oellien, Wolf-Dietrich Ihlenfeldt, Klaus Engel, Thomas Ertl

Computer-Chemie-Centrum, Institute of Organic Chemistry, University of Erlangen-Nürnberg, Nägelsbachstr. 25, D-91052 Erlangen
WWW: <http://www2.chemie.uni-erlangen.de/ChemVis/>, email: {Frank.Oellien, Ihlenfeldt}@chemie.uni-erlangen.de

Visualization and Interactive Systems, Institute of Computer Science, University of Stuttgart, Breitwiesenstr. 20-22, D-70656 Stuttgart
WWW: <http://wwwvis.informatik.uni-stuttgart.de>, email: {Engel, Ertl}@informatik.uni-stuttgart.de



Visualization and Interactive
Systems Group

Abstract

Today, chemical companies are routinely employing synthesis technologies such as High-Throughput-Screening, Combinatorial Synthesis and Parallel Synthesis that generate terabytes of data per year. These datasets are highly relevant for chemists, because they potentially deliver insights into chemical and biochemical trends and principles and could lead to faster development and higher numbers of drug candidates. However, the development of data mining and visualization tools that are capable of analyzing these large amounts of data appears not to have been able to keep pace with the dramatic increase in size of these datasets. Ultimately, this situation has become one of the most critical bottlenecks in chemical R&D today.

We present novel methods for the interactive graphical visualization and mining of large multi-dimensional and multi-variate datasets. The capabilities of the presented applications are demonstrated with a large reaction database (courtesy of ChemCodes Inc.).

Technique

Glyph-based InfVis Approach

Chemical datasets have not only increased in size, they also have become multi-dimensional. For the visualization of such multi-dimensional data from chemistry, we have implemented an Internet-enabled Glyph-based chemistry data visualization tool that utilizes 3D hardware capabilities of modern desktop clients. The Java/Java3D-based application can be run as stand-alone program or as an applet embedded in a web page and is platform-independent. A major design goal was straightforward usability of the applet. The application does not require expert knowledge and is intended to be usable by laboratory chemists. By using Glyph technology, an user may map up to six data dimensions onto the three orthogonal axes and to shape, size and color of the displayed scene objects. Additional dimensions can easily be added by dynamic query devices.

The application has three subpanels:

- the 3D render panel
- the tool panel
- and the control panel.

These subpanels are connected via split plane bars that allow panel resizing and hiding.

Java3D Render Panel
Within this panel the datapoints are displayed. Java3D technology allows fast and high-quality rendering of the data. The user can examine the data from different viewpoints or can navigate (zoom, translation, rotation) through the 3D world. By using axis range sliders, parts of the data can be temporarily hidden. Currently, the user can choose between the *glyph* and the 3D *barchart* display style. *Surface* and *point cloud* styles will be added in the next release.

Tool Panel
This panel contains several tools like dynamic query devices (sliders, range sliders, checkboxes) and selection tools that allow interactive mining and analysis of datasets. Resizable and movable, semitransparent selection boxes facilitate the selection of interesting data subsets. Interactive filters are used to mine and analyze the datasets in realtime.

Control Panel
This panel controls the data input and display options for data visualization. It contains interfaces that link to databases via JDBC and is thus able to retrieve nearly arbitrary data from remote database tables. Via the mapping interface, the user may map the data dimensions onto axes and objects in the 3D scene. In addition, it provides various control options to change display attributes.

Interactive mining by means of voice commands

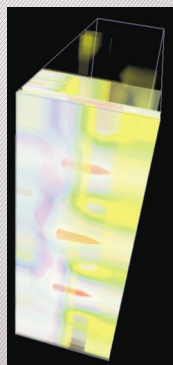
The use of dynamic query devices and hardware-accelerated 3D rendering makes it simple and straightforward to interactively and graphically mine large datasets. However, we believe that an agent-driven mining by means of a voice command recognition interface should increase the quality of user interaction measurably. It also will simplify the use of the application. We are already working on the integration of the Java Speech API (JSAPI) into the tool.

Volume-based InfVis Approach

The Glyph-based prototype is able to handle from a few hundreds up to a few thousand data points. For the visualization of larger datasets, which are certainly not uncommon, alternative approaches such as volumetric methods are required. We have developed a novel texture-based volume rendering approach which is especially suited for low- to midsize resolution volume data with non-linear transfer functions. This method, called *pre-integrated volume rendering*, provides a high image quality and good interactivity due to high frame rates.

This technique is capable of visualizing several million datapoints. Three data dimensions are mapped to the orthogonal axes and a fourth dimension is represented by the RGB value of the voxels. Additional dimensions can be added by combining dynamic query devices and specific transfer functions.

We are planning to combine the advantages of the Glyph-based and texture-based approaches in future program releases. The texture-based technique will be used for an overview. For detailed analysis of selected subsets, the applet will switch to the Glyph mode.



Examples

Reaction Optimization

Reaction optimization is a major problem for chemical companies. Higher product yields and less side products result in lower costs. By careful studying of reaction data it is possible to find suitable reaction conditions. However, this information is often difficult to obtain, because reaction data will only be published for successful reactions and with unreliable yields. Experiments that did not yield the expected results will usually not find their way into publications. The documented reaction space has large holes with no information. ChemCodes Inc. has begun to build a reaction database that will attempt to solve this problem. Ultimately, it will cover a large fraction of the common reaction space and include valuable data about unsuccessful experiments. We used a database subset and our tool to analyze reaction conditions for aldole condensation. Here, Tentagel is a better resin than polystyrene and the number of side products increases with temperature. The best results are achieved at 23°C on Tentagel and potassium hydroxide as activating base.

Reaction Database Subset with 66 Datapoints

| X-Axis: | Reaction category | Rating | Description |
|---------|--------------------------------------|--------|-------------|
| 1 | Target product only | | |
| 2 | Target product and starting material | | |
| 4 | Starting material only | | |
| 7 | No data | | |

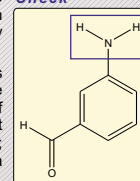
Y-Axis: Solvents
Z-Axis: Resin: Polystyrene (PS), Tentagel (TG)
Object size: Yield
Object color: Reaction category
Object shape: Temperature (25 °C - sphere, 60 °C - box)

Reaction Planning

Another common problem in chemical laboratories is the search for reaction conditions that allow the selective execution of a specific reaction path from among several possibilities. Precursors that contain a reactive functional group for a planned reaction often contain other reactive functional groups that may interfere under many of the possible reaction conditions. In the past, chemists solved this problem by intuition and fuzzy knowledge and generally by *trial and error*.

We have used a ChemCodes reaction database subset and our InfVis application to find reaction conditions that allow the selective attack on the phenyl amine group of 3-aminobenzaldehyde. By the application of toolbox filters, we were able to isolate several reaction conditions that let the amino group react, but not the aromatic aldehyde. (left: full dataset; right: filtered dataset; solvents: THF, pyridine or DMF; reagent: ammonia in EtOH; quencher: phenylthiourea)

Functional Group Compatibility Check



Database Subset with 792 Datapoints

| X-Axis: | Number of products |
|---------------|--------------------|
| Y-Axis: | Solvents |
| Z-Axis: | Functional Group |
| Object color: | Functional Group |

References

- Ihlenfeldt, W.D.; Oellien, F.; Engel, K.; Ertl, T. "Multi-Variate Interactive Visualization of Data from Digital Laboratory Notebooks", *ECDL 2001, Workshop "Generalized Documents: A key challenge in digital library research and development"*, September 8th, 2001, Darmstadt.
- ChemCodes Inc., Durham, NC, USA, <http://www.chemcodes.com>

ChemVis Homepage:
<http://www2.chemie.uni-erlangen.de/ChemVis/>