

---

Department of Computer Science  
University of Nevada, Reno

**A Software Tool for Data Visualization:  
Text Mining and Clustering**

Team 3  
Brett Boge  
Danielle Idland  
Nolan Fleming  
Russell Hardie

Sergiu Dascalu, *Professor*

Jeff Elpern, *Advisor*

*Tuesday, February 28, 2006*

---

## Abstract

---

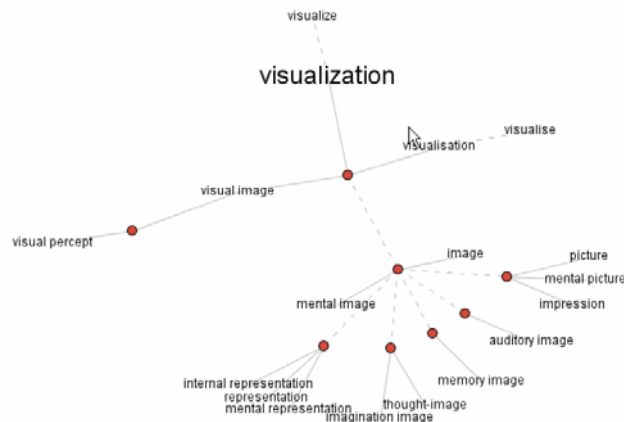
Within the expanding field of software engineering, there is extensive growth in the amount of data being produced. As the quantity of data increases beyond what is practical for users to analyze independently, tools are needed to assist with this task. Data visualization tools are best suited to this task.

One arena for such a project is the internet, which holds a massive amount of data. Many different kinds of websites, whether they are online libraries, professional and academic journals, or even business or personal finance sites, have a lot of data to present to their users. One way to allow users to collect the data best suited to them is to use a web-crawler; to further develop the application into something truly useful, data visualization tools can be included. A particularly useful tool for searching through data yielded by the web-crawler is to look at a cluster analysis (a diagram where clusters of like items are linked together).

## Main Goals

---

The goal for our Senior Project is to create a system of data visualization based somewhat on the field of “Text Data Mining” which is currently a focus in such groups as SQI Inc. and our very own University of Nevada. The idea of data mining is to take large amount of text and extract useful data from it, converting the data into an easily comprehensible graphical analysis. We would accomplish by constructing a visualization system to display data based on clusters of likeness and relationships; with a structure similar to the example from the Visual Thesaurus shown below:



## Users

---

“Users” is a term that will be used slightly more broadly for this project than for many other projects. Because our project is essentially a visualizer for large data sets, we suppose that it would be very useful in office presentations where many people would have to read and interpret the final product. In that sense, the “users” of this program would not only be those people who have direct contact with the computer and the software therein, but the group of users would include all people present for the presentation. However, our project will not be limited to this use. Other users could include any individual who wishes to write a report or merely see their information in a different way.

## **Functionality**

---

To summarize the inner workings of our project, it would take in a text file of a predetermined format and parse through it, keeping track of certain statistics. These statistics could include such important factors as time duration, density of samples, and relationships between clusters of data. We could also parse through large text files (or sets of text files) searching for specific words or phrases. After parsing through the documents, our program would output a visual schematic or graph appropriate to the subject matter and desired information interpolated from them. A friendly graphical user interface would enable users to choose specific settings pertaining to the input file as well as the output visualization.

## **Challenges**

---

Of course, we expect many challenges, all of which we will face head on. The first and foremost of our challenges ahead pertains to the main purpose of the project itself. We do not yet know exactly how we will go about parsing through a variety of text files and then output a fitting display for all the information, but that should only remain a problem for a short while. The next biggest challenge lies with us, the people. Working with a group of people always has some risk attached to it. The trick is to decrease and manage that risk to an acceptable amount. Once we establish specific roles for every person, that problem should be all but eliminated. Other problems are currently unknown, but they will surely be taken in stride.

Our project is intended to run solely on personal computers using a combination of languages including C++ and Python.

## **Professional Growth**

---

This project helps the professional growth of all team members as research project to collaborate with our colleagues, and the opportunity to produce quality work. It also allows the opportunity to work with new or unfamiliar languages, to work with data manipulation and graphics, and to further develop existing skills.

## **Project Timeline**

---

Dates the due dates for each individual component and are subject to change. Priorities are ranked from high to medium to low, with high priority referring to items the project must have functional, medium priority referring to items the project should have functional to be a complete project, and low priority being items that add to the quality of the project, but that the project is completed without.

### **Phases**

*Design – to be completed by March 10<sup>th</sup>*

Design of all parts.

Deliverables:

Product specifications.

Project website.

*Implementation – to be completed by April 14<sup>th</sup>*

Coding of project.

Deliverables:

Integration of web-crawler code: high (3/24)

2D display: high (3/31)

Neighborhood view of nodes: medium (4/7)

Extended neighborhood view: low (4/7)

Change central node: medium (3/31)

3D display: medium (4/7)

Switching between 2D-3D functionality: low (4/14)

*Testing - to be completed by April 28<sup>th</sup>*

Testing of final project, to include performance and stress testing.

Deliverable:

Results of tests.

## About the Team

---

**Brett Boge:** Brett is a senior at UNR, and an intern at IGT. He has a broad range of talents, with experience in C++, Java, Python, and graphics programming. He likes solving problems and having new challenges to face. In his free time he likes sliding down snowy slopes with a piece of plastic strapped to his feet.

**Nolan Fleming:** In between being prepared for zombie and/or ninja attacks, Nolan attempts to take every class possible at UNR prior to graduation,.. or at least in the computer science department. With experience in C++, Java, assorted low-level languages, and graphics, Nolan is an asset of our team focused on coding and getting stuff done, no matter what it takes.

**Russell Hardie:** With experience in C/C++, Java, and OpenGL, Russell is an asset to our team focused on graphics.

**Danielle Idland:** A CS major by day, federal employee by really early morning, Danielle splits her time between work, school, and the occasional nap. With experience in C/C++, HTML, and some brief time spent dabbling with MySQL, Danielle is focused on the code-oriented and design parts of the project.

## About our Advisors

---

**Jeff Elpern:** Provided code for web-crawler and original inspiration for the idea. His role in the project will be to help keep the team on track and to make sure that it doesn't deviate from the customer standards, as well as to provide additional inspiration for future development.

**Sergiu Dascalu:** Professor of Software Engineering at the University of Nevada, Reno. His current research interests include software engineering (in particular, languages and techniques for software specification and environments for software development), human-computer interaction, formal methods, and real-time systems.

## References

---

- A New Generation of Data Mining Technologies*  
by ANGOSS Software Corporation  
<[http://www.dmreview.com/whitepaper/paper\\_sub.cfm?whitepaperId=10063](http://www.dmreview.com/whitepaper/paper_sub.cfm?whitepaperId=10063)>
- Hearst, Marti A., Text Data Mining: Issues, Techniques, and the Relationship to Information Access  
July, 1997. <<http://www.sims.berkeley.edu/~hearst/talks/dm-talk/>>
- Text Mining with Information Extraction*  
Un Yong Nahm Ph.D. Thesis, Department of Computer Sciences, University of Texas at Austin, August 2004.  
<<http://www.cs.utexas.edu/users/ml/papers/discotex-dissertation-04.pdf>>
- Fayyad, Usama; Grinstein, Georges; Wierse, Andreas, Information Visualization in Data Mining and Knowledge Discovery.  
San Francisco: Morgan Kaufman, 2001.  
<<http://books.elsevier.com/mk/?isbn=1558606890>>
- Keim, D.A., Information visualization and visual data mining. Visualization and Computer Graphics,  
IEEE Transactions on. Volume 8, Issue 1, Jan.-March 2002 Page(s):1 - 8

## Useful Websites

- Model & Mine, Dorian Pyle*  
<http://www.modelandmine.com/index.htm>
- Text-Mining and Data-Mining Software*  
<http://www.data-miner.com/>
- Data Mining and Discovery*  
<http://www.aaai.org/AITopics/html/mining.html>
- Text Mining Links and Resources*  
<http://www.text-mining.org/>  
<http://www.textminingnews.com/>  
<http://www.textanalysis.info/content.htm>