

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

Danielle Idland		universitychica@gmail.com
Brett Boge		brettb@brettb.net
Nolan Fleming		fleming_nolan@yahoo.com
Russell Hardie		rhardie126@msn.com
		and
Sergiu Dascalu, PhD		dascalus@cse.unr.edu
Jeff Elpern		jeff.elpern@sqi-inc.com

Department of Computer Science and Engineering  
University of Nevada-Reno  
Mail Stop 171  
Reno NV, 89557

### **1. Synopsis**

The software for a data visualization tool presented in this paper allows for the ability to readily identify the most useful and relevant information presented to researchers, students, and other end users, by the data and text mining queries performed by a web crawler. It includes explanations of data mining, text mining, and web crawlers.

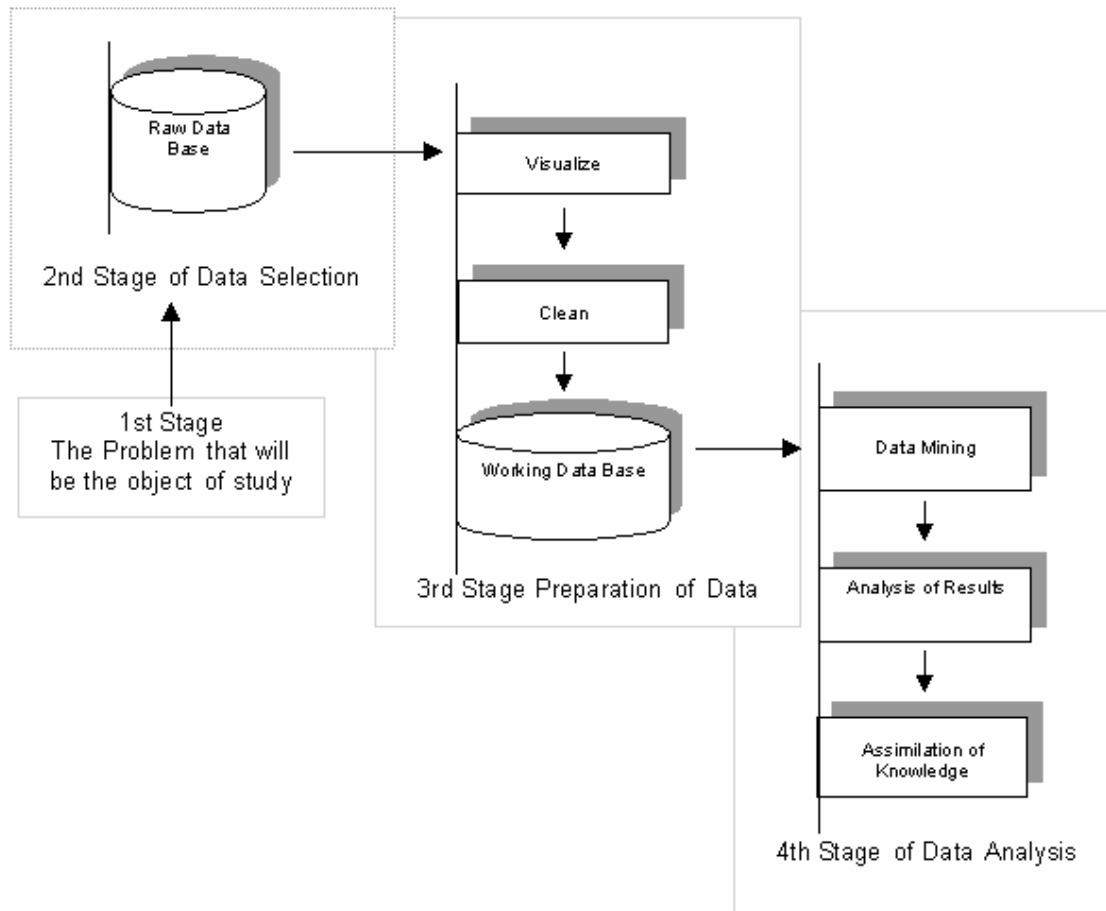
### **2. Introduction**

As the computing industry grows, there is an increasing amount of data available to both computer users and their machines. The problem that occurs is that there is no reasonable means for a computer or person to peruse all this data in a reasonable amount of time, and searching through a mass of irrelevant data to receive useful information wastes valuable assets.

Search engines on the Internet are an answer to this problem. A person can utilize a search engine and find, in most cases, data that is useful for their needs in a reasonable amount of time. The problem with many search engines is that relevance is hard to determine quickly and still more valuable assets are wasted while trying to find the most relevant data.

Search engines work via the use of web crawlers. Web crawlers are automated programs that browse through the web methodically. They are predominately used to obtain data for search engines to use as a catalog. The web pages are downloaded and can be accessed at a faster rate when a query is put into the search engine. Web crawlers also can be used for many other functions including maintenance related tasks such as validating links on a website. As such, web crawlers are useful tools that save a lot of time and effort for end users.

Web crawlers can be used for tasks such as data and text mining [1, 2, 3]. Data mining, which is also known as Knowledge-Discovery in Databases (KDD), is processing large volumes of data for noticeable patterns. These patterns are then used for data analysis. Data mining provides users with data that they can make sense of and use. The patterns provided by data mining are of significantly more use to researchers than the original data en masse would have been. The patterns can also save a lot of valuable time for end users and help them to realize the true relevance of the data [1, 3].



*Figure 1: Data Mining Process [6]*

In a similar fashion, text mining searches large amounts of text for useful information. It is the text equivalent of data mining, and is also known as Knowledge-Discovery in Text (KDT) and intelligent text analysis. Text mining is useful as more than 80% of data is stored as text. It is also thought to have both commercial and research applications because of the high volume of text-based documentation.

The main problem with text and data mining is the same problem that search engines have: an overabundance of information without the ability to readily distinguish how useful information is. The user is still forced to spend time making sense of the results

they've received. Visual interpretation of data, rather than textual, is one of the best approaches to this problem.

### Description of the Research

Our team is developing a software visualization tool that allows users to visually determine the relevance of data they are analyzing. It helps users to make and understand connections from their data, and to easily identify outliers in their data set.

The software sees each piece of data as a different node. Each node has an identity pertaining to its vector position in relation to the overall query and its connection to similar (or neighboring) nodes. The nodes are visually represented in both two-dimensional and three-dimensional graphs that can be scanned by the user. Users can also choose which node is the centrod, or center of the graph, and is thus the source of the query and visualization [3, 5].

In addition, a clustering method is used to distinguish relationships in between the data in various vectors. Like items are clustered into groups together and separated linearly from similar items that are outside of the cluster. This allows for ready identification of patterns within the data set.

### Results

Data is portrayed in a fashion that is easy to understand and manipulate. The data can be scanned by changing any node to the centrod and then evaluating the new set of neighboring nodes and their relationship to the central node [3]. Patterns are readily identifiable to users and can be evaluated by changing the centrod.

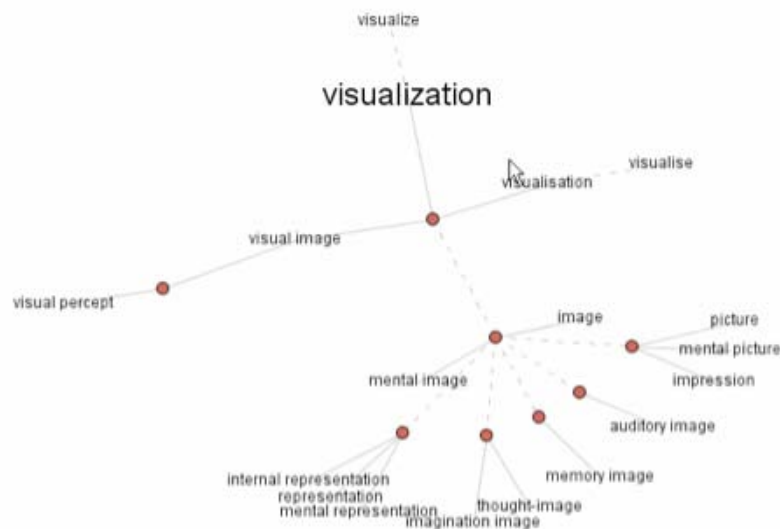


Figure 2: Example Image from Visual Thesaurus [4]

This example demonstrates the linear relationship representation of the word “visualization”. Relationships include those between similar words (visualization, visualize, and visual image) and those of similar concepts (visualization, image, and mental image), as well as the neighboring nodes of everything having an immediate relationship with “visualization”. It also shows the example of clustering by presenting the various types of images with sub-clusters that are more specific.

## Conclusions

With the constant technological expansion comes a great need for the ability to quickly and accurately process massive amounts of data. A software visualization tool is an excellent resource for users who need the functionality of data processing without the burden of a time consuming methodology. A graphical representation is the most effective visualization in that it allows users to see instantaneously recognize the patterns and relevance they were for which they were searching.

## References

- [1] Marti A. Hearst. Untangling Text Data Mining. School of Information and Systems, University of California, Berkeley, 1999. <<http://www.sims.berkeley.edu/~hearst/papers/ac199/ac199-tdm.html>>
- [2] Shrikanth Shankar and George Karypis, Weight adjustment schemes for a centroid based classifier. University of Minnesota, Department of Computer Science, Technical Report: TR 00-035 <<http://www.sqi-inc.com/twiki/pub/Univ/UNRTextMiningWorksheet/WeightAdjustedSchemesForCantriodBasedClassifier.pdf>>
- [3] T. Nasukawa and T. Nagano, Text analysis and knowledge mining system , IBM Systems Journal, VOL 40, NO 4, 2001 <<http://www.sqi-inc.com/twiki/pub/Univ/UNRTextMiningWorksheet/Textanalysisandknowledgeminingsytem.pdf>>
- [4] Thinkmap Visual Thesaurus. <<http://www.visualthesaurus.com>>
- [5] Wikipedia, The Free Encyclopedia. <<http://en.wikipedia.org/wiki/>>
- [6] Tarapanoff, Kira, Luc Quoniam, Rogério Henrique de Araújo Júnior, and Lillian Alvares. Intelligence obtained by applying data mining to a database of French theses on the subject of Brazil. Information Research, Vol. 7 No. 1, October 2001. <<http://informationr.net/ir/7-1/paper117.html>>