Efficient and interactive 3D visualization of large mined datasets

### Undergraduate Honours Project (COMP 4520) proposal prepared by Levko Ivanchuk, 7670173

## Project outline

Today, Big Data is a term commonly used to describe the vast amount of information harvested by big corporations & other institutions every day. This information is then used to provide deeper insights and strengthen the decisions made by such organizations. More and more companies employ different tools and techniques from various areas of data mining to extract non-trivial and potentially useful knowledge from their data. Visualizing extracted knowledge in a simple and intuitive manner remains a difficult task, mainly because of the volume of the information, its variety and its velocity. In this project, we aim to develop a visualization system that takes advantage of three-dimensional space and interactivity to improve comprehension and understanding of big datasets. By developing an intuitive, fast and easy-to-use system, we plan to improve ones comprehension and understanding of data.

Our aim is also to develop a visualization that is both quick to interact with and requires minimum training to understand and use.

## Problem

Visualization of data mining results is perhaps more important than actually data mining itself. Mined data is of no use if people analyzing and interacting with it can not quickly understand what is going on and what the data is “telling” them. Researchers in areas of both data mining and visual analytics have looked at big data visualization for many years.

The effect – users will be able to easily see how frequent itemsets relate to each other, what is their actual frequency. A user that has never worked with a given system must be able to learn to use it with minimum training and be able to read the mining results right away.

1) what’s the effect of a novel technology? 2) what’s the frequency or rate of a factor like the adoption of a technology? 3) presenting a novel technology and devise the possible scenarios in which it could be used presenting initial user feedback or other advantages like simplicity or robustness over previous options. 4) what’s the cost and risks factors associated with a certain novel technology?

Outline of the project's scope (abstract)

A background of the problem (literature review)

Research direction (proposed methodology)

Facilities (computing or other) required to do the research

An anticipated length of time requirement

Anticipated outcome

A list of relevant references

What have I learned throughout the course project

* A lot of data takes a lot of time & memory to process
* A modification to the visualization shape, at least in primitive implementation, requires a lot of recalculation (sometimes redundant recalculation)
* There are many ways to store, parse and display billions of data points very efficiently - it is just that we didn’t use them in our project
* Clustering on large data can be done more efficiently, however, due to the previously mentioned problems, it is hard to test & even think about

Solutions

At this point in time there are 3 major areas that can and should be improved to make the whole visualization better.

* Data processing and shape generation
* Visualization load times & usability in terms of performance
* Interactivity of the visualization + the amount of information one can find while viewing

These three points are very vague and general, however, they should define the research scope . A different approach to processing and displaying the visualization is proposed at the next page, as well as some additional features to the interactivity part.

To achieve the following goals (although they are too abstract and vague) I am proposing the following approach, illustrated here:

VIEWER

* accepts an meta-data augmented file + Polygon file for the shape itself
* allows user to interact with the visualization
* is optimized for viewing, not requiring the original data
* users are given extra features made possible by the metadata

PARSER

- goes through the data

* provides non-interactive preview of the shape and appearance of the visualization
* allows designer to change some parameters & preview the final visualization
* generates an meta-data augmented Polygon file
* lots of parallel processing & optimizations

.PLY file + meta data

pre-processing

I propose a some sort of an add-on for the mining process itself. By wiring it to the mining process, we can simultaneously generate the visualization data. Later, after the mining process is finished, one can preview the visualization as a screenshot, while also adjusting the parameters of the visualization (eg. shape height, frequency line threshold, clustering boundaries per level, etc.) and getting an almost instant rendering of the final visualization. The goal is to support AT LEAST 4 million records and the higher - the better. To do that we will employ CUDA parallel processing on the GPU, as most of the operations are highly adaptable to parallel processing.

additional interactivity

This comes in the viewer part. Users should be allowed to filter by level, take a look at one level in isolation, drill down into the cluster, select an individual item set even if the number of item sets is very large. Also, a search feature will be useful. Again, some parallel processing might be needed here, but what we are sure of is that we have the technology to draw and render billions of data points in 3D. The rest we will have to research & develop.