ConeViz: Efficient and interactive 3D visualization of frequent patterns

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

## Project outline

Today, companies gather large amounts of data. They use technique called ‘frequent pattern mining’ to discover implicit, previously unknown and potentially useful knowledge from their data. For example, in a supermarket, frequent pattern mining might discover that bread is frequently purchased with butter, or milk is often purchased with cookies, etc. This information is important to businesses, as it helps them to understand the behavior of their customers, and thus choose a more efficient marketing strategy, or take some other actions to increase their revenue. Items that were frequently purchased together are called frequent patterns. Such patterns are presented to the user in textual form and are often called itemsets. Please see Figure 1 for an example of such output.

Bread (37)

Milk (146)

Milk (146) Bread (10)

Milk (146) Cookies (10)

Milk (146) Cookies (10) Chocolate bar (10)

Milk (146) Corn flakes (10)

Figure 1 – An example of FREQUENT PATTERN MINING OUTPUT  
An itemset is represented as a single line. It contains a list of item names, followed by their frequency (i.e. number of purchases in some timespan). In this instance, we can see that bread was purchased 37 times, whereas milk together with bread was purchased 10 times, etc.

Navigating through a large list of itemsets can sometimes be difficult. Often these lists would contain a couple of million records, which renders the task of discovering interesting information even more complex and time consuming for the end user. In such cases, it is common to apply some visualization technique to the data. By representing the same dataset in a different way, users can benefit from better understanding of their data, and as a result, better decisions.

In this project, we aim to develop a visualization that takes advantage of three-dimensional space and interactivity to improve comprehension and understanding of frequent pattern datasets, such as one presented in the Figure 1. Our main contribution is the proposal and development of a 3 dimensional frequent pattern visualizer that can be scaled to large datasets and remain highly interactive, while giving the user feedback about the frequency of the itemsets. The primary focus of the visualization is to help users easily discover all the temsets that contain a particular item.

## Related work

Visualization of data mining results is as important as the data mining itself. Mined data is of no use if people analyzing and interacting with it can not quickly understand 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. Some examples include FIsViz [1], FpVAT[2], PowerSetViewer[3], Yang’s system [4][5]. We briefly discuss some of them in the remainder of this section.

**FpVAT**

FpVAT is one of the recently developed frequent pattern visualizers. Consisting of two distinct modules, this system allows users to get an overview of the massive datasets, so that they can derive insights from it. Another module allows users to perform analytical reasoning via interactive visual interfaces to assist with detecting the expected frequent patterns and discovery of the unexpected ones.

The system uses a polyline method to connect the itemsets displayed on a 2 dimensional plane. As datasets grow larger, the number of lines that FpVAT displays will grow large, which in turn may create distractions to the user. In addition to that, it is not immediately obvious how to read the information that the system displays. We plan to avoid the clutter by using a data visualization technique called Edge Bundling, or grouping many adjacent edges into one. In addition, ConeViz allows user to view the exact frequency of any itemset.

**PowerSetViewer**PowerSetViewer is a frequent patter visualizer that groups all patterns together based on cardinality and presents them in a two dimensional grid. It applies color coding to the background to indicate the cardinality of the itemsets. Whilst guaranteeing visibility, this visualization system also groups multiple patterns into the same square, which makes it hard to distinguish individual itemsets in the visualization. In addition, the system does not show the exact frequency of any given itemset. Our proposed system will allow users to see the exact frequency of any itemset. We also plan to apply grouping to deal with sheer size of the data; however, users will be able to drill down into groups or clusters, such that individual itemsets are accessible.

**Yang’s system**

Designed to visualize association rules[[1]](#footnote-1), but can also be used to visualize frequent patters, Yang’s system plots domain items[[2]](#footnote-2) in a two-dimensional space consisting of many vertical axes. Itemsets are then represented as connections between these domain items, where the thickness of the line connecting them depends on the frequency of the itemset. This can be considered as a potential problem, because the thickness of the line gives users only an approximate estimation about the actual frequency of the itemset. In addition, Yang’s system lacks support of many interactive features that are required in large datasets. We plan to address these issues by always displaying the exact frequency of a currently selected itemset. In addition, we propose a search feature for quick lookup of a domain item(s) or an itemset.

Our proposal differs from previous approaches by 1) using 3D visualization, 2) allowing for many interactive features, such as searching and exact frequency display, 3) allowing users to select a single itemset or a domain item and interactively ‘see’ all the related or connected itemsets.

## Research direction

In our work with investigate a 3D visualization, called ConeViz, for data mining results. We believe that this approach can improve both learnability and scalability of the visualization by allowing for more interaction techniques (such as zooming in/out, rotating, flipping and spinning of the 3D model) and by simply allowing for more space in a 3D visualization.

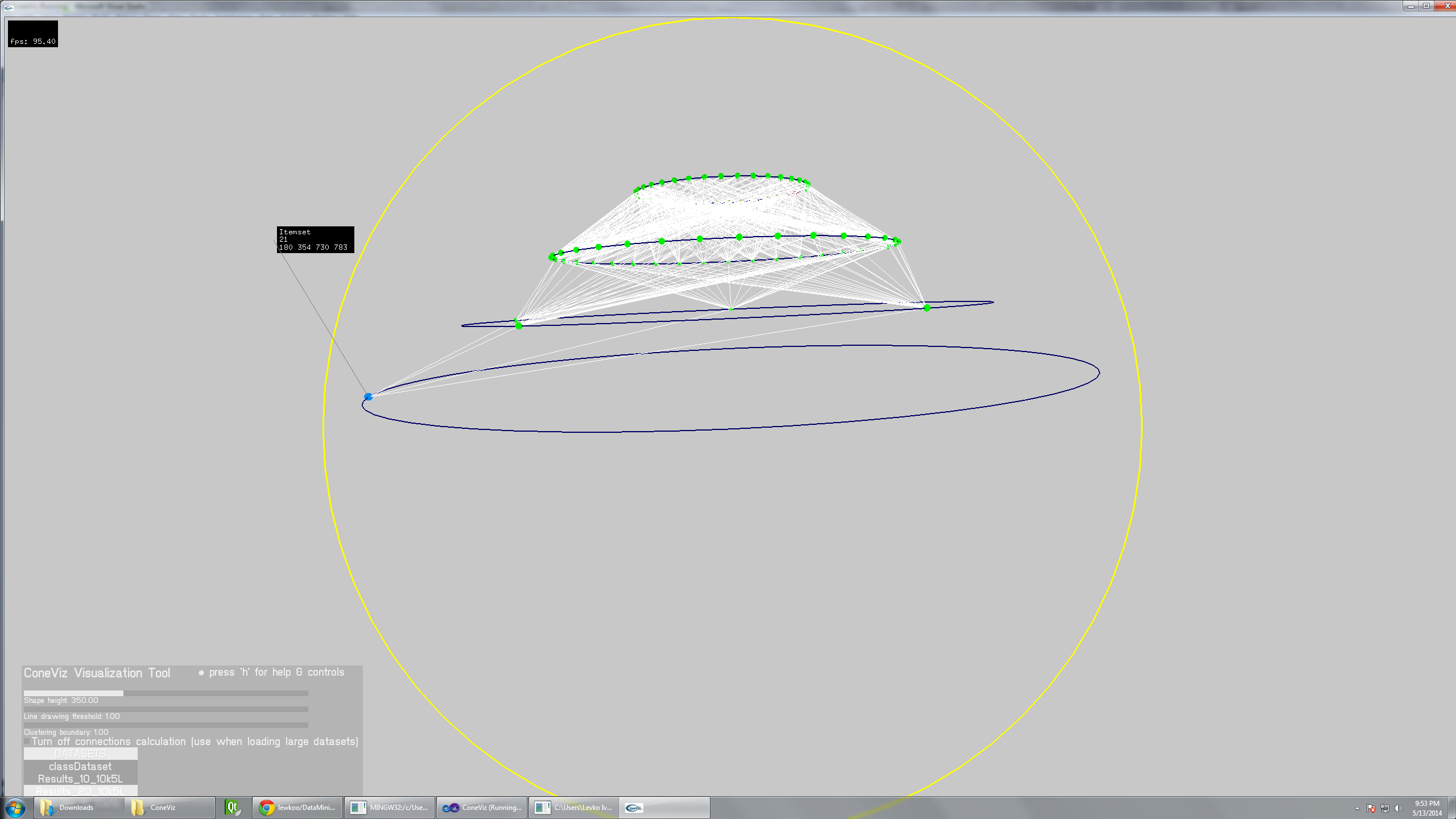
****

Figure 2. Initial Prototype

Figure 2 shows a preliminary design of the visualization. Itemsets are represented as colored spheres, placed on circles (i.e. levels). White lines represent the connections between the itemsets. Users also get a message box displayed next to their mouse cursor indicating the information about the selected itemset. By selecting an itemset, users are able to see only connections in the tree leading to that itemset, with other distractions removed. This leads to better pattern recognition and discovery, as users are able to search for domain items or itemsets they are interested in, and quickly find all the related items in the frequent pattern dataset. The visualization will also hide all irrelevant data points, so that users can focus solemnly on the data that interests them. We believe that this approach will allow users to quickly discover patterns that they are interested in, while hiding the irrelevant data.

We see three main parts to this project:

* Data processing, shape generation, performance analysis.
* Interactivity of the visualization + the amount of information one can find while viewing
* Visualization usability

Some of the difficulties while working with large data sets involve processing times and memory consumption. Therefore, we propose to split the processing and shape generation from the viewing and interaction, as shown on Figure 3.

DATA PARSER

- goes through the data

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

VIEWER

* accepts a 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 (sorting, searching, etc)

visualization files + meta data

Figure 3 – initial project structure

We see the following benefits in adopting this structure. Due to the high number of records in the itemset databases, it is highly likely that processing & shape generation will take some time. Therefore, it might be beneficial to be able to generate the visualization once and then view it later, instead of generating the visualization from scratch at every view request. In addition, this allows us to experiment with visualization itself on the data parser level, without changing the viewer. For instance, if later in the project we decide to take a different approach with our visualization, we would only need to modify the data parser to generate the updated visualization. The viewer will remain unchanged.   
  
DATA PARSER - We propose an add-on for the mining process itself. By wiring it to the mining process, we can simultaneously generate the visualization data. This approach, however, is highly complex, so we might take a different approach of simply parsing the already generated frequent pattern data files, as seen in Figure 1. 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 or higher. To do that we will employ parallel processing on the GPU (either shader or CUDA based), as most of the operations are highly adaptable to parallel processing.

Viewer will take the generated files from the parser and efficiently display them. Users will be allowed to interact with the shape, rotate it in any direction, as well as zoom in and zoom out. In addition, users will be able to select a given itemset and see all the connections leading to that particular itemset, Users will also see the exact frequency of the itemset, as well as all items that are in that selected set. Users should be allowed to filter by level, take a look at one level in isolation, drill down into the cluster, select an individual itemset even if the number of itemsets is very large. Also, a search feature will be useful.

Additional directions of this project could include exploration of different 3D shapes that could potentially be useful at displaying frequent itemset mining results. Since there is a substantial amount of research on this topic in the areas of visual analytics, we might spend a bit more time reviewing the existing literature to find a suitable shape. In addition, we can cooperate with the HCI Lab at University of Manitoba in case we need to develop a better shape or paradigm to visualize and display frequent itemset mining results in 3D.

It would also be interesting to explore Unity for this project. Unity is a 3D graphics development suite that is becoming very popular and is increasingly being adopted by research community. It allows for quick and easy scripting of 3D shape & scene generation, as well as interaction with it. Using Unity will definitely add some novelty to this project and make it up-to-date with current 3D development techniques. However, using Unity might also be a limiting factor, as it is currently unknown to us how restricted the system it. Therefore, adopting Unity should be considered with caution.

## Facilities

Since both parser & viewer will be aimed at a low-performance computer, there is no extra computing facility required for this project. Frameworks and utilities that will be used must be open-source. The project should remain cross-platform, regardless of the chosen direction (i.e. Unity development or some other OpenGL frameworks).

## Anticipated length

Time wise, this project highly depends on the number of interactivity features that the system will support. Therefore, it will be important to develop an efficient and adaptable data parser, such that it can be modified to produce more metadata for the viewer, if needed. Since we already have some preliminary code for both data parser and the visualizer, and know where its weaknesses are, data parsing step should require no more than a month or 1.5 month of development. This leaves another 2 – 1.5 month to develop the viewer, evaluate it and wrap up the project by perhaps writing a short paper. Of course, these length estimations should be changed if we take the decision to use Unity for this project.

## Anticipated outcome

The main goal is to build a complete frequent pattern visualization system and evaluate if 3D visualization is of any potential benefit for visualizing this particular type of data over existing 2D visualizations. In addition, this project should produce suggestions about different data representations that can be used to visualize frequent pattern mining results in 3D.

## References

[1] C. K.-S. Leung, P. P. Irani, and C. L. Carmichael, “FIsViz: A Frequent Itemset Visualizer,” in *Advances in Knowledge Discovery and Data Mining*, T. Washio, E. Suzuki, K. M. Ting, and A. Inokuchi, Eds. Springer Berlin Heidelberg, 2008, pp. 644–652.

[2] C. K.-S. Leung and C. L. Carmichael, “FpViz: A Visualizer for Frequent Pattern Mining,” in *Proceedings of the ACM SIGKDD Workshop on Visual Analytics and Knowledge Discovery: Integrating Automated Analysis with Interactive Exploration*, New York, NY, USA, 2009, pp. 30–39.

[3] Q. Kong, “Visual mining of powersets with large alphabets,” 2006.

[4] L. Yang, “Pruning and Visualizing Generalized Association Rules in Parallel Coordinates,” *IEEE Trans. Knowl. Data Eng.*, vol. 17, no. 1, pp. 60–70, 2005.

[5] J. Yuan, Y. Wu, and M. Yang, “From Frequent Itemsets to Semantically Meaningful Visual Patterns,” in *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, New York, NY, USA, 2007, pp. 864–873.

1. For instance, if a customer purchases milk and bread, association rule mining might discover that in X% of cases a given customer also purchased butter. This information can be useful for creating effective promotions and sales. [↑](#footnote-ref-1)
2. Domain items are unique items in a single dataset. For instance, in a dataset of a supermarket, unique items such as ‘Bread’, ‘Milk’, etc. are considered to be in a list of domain items. Any combinations on these are not considered to be in a list of domain items. [↑](#footnote-ref-2)