Subject name: Visual Analytics

Lecturer: Zhonglin Qu

Student Name: Xuan Bach Tran

Student ID: 22027453

Assignment 1

DECLARATION

I hold a copy of this assignment that I can produce if the original is lost or damaged.

I hereby certify that no part of this assignment/product has been copied from any other student's work or from any other source except where due acknowledgement is made in the assignment.

No part of this assignment/product has been written/produced for me by another person except where such collaboration has been authorised by the subject lecturer/tutor concerned.

Introduction

In this digitised age, the increasing amount of data nowadays may vary in different forms, sizes, and types; they can be structured or unstructured. Henceforth, there has long risen the demand to manipulate, store, and process them in ways that help human visualise. Zhang and his co-authors stated that the large volumes of data fuelled extensive data analysis research (Zhang et al. 2011). Along with data analysing, there are visualisation tools available that help present the analytics visually. It is proven by many studies and research that effective data visualization enhances communication and knowledge discovery, making it an indispensable aspect of data analysis (Munzner, 2014).

Data visualisation is regarded as communication and conveyance of information through visual elements such charts, graphs, and maps. Tools such as Tableau, PowerBI, Gephi, etc. were created to support that task; they transform raw data sets to understandable story, users and readers can get a grasp of the big picture, identify trends and patterns that support their decision-making process at organisation level (Fews, 2009).

Literature Review

Illustration methods take different forms and shapes, each will require different skills and knowledge to properly visualise them. According to Nguyen & Huang (2005), data can be illustrated via connection approach, enclosure approach, or hybrid approach. The connection approach shows the relationships in the hierarchies using node-link diagram, whereas the enclosure approach optimises the display space by leveraging enclosed shapes to demonstrate hierarchical levels. The two methods will be discussed in this report with Degree-of-Interest Tree and Treemap, alongside with a different approach namely Choropleth map, which was applied during the process of analysing the datasets.

Degree-of-Interest Tree (DOITree):

The DOITree was pioneered by Jefferey Heer and Stuart. K. Card as a solution for the slow growth of display size and resolution, despite the exponential multiplication of dataset variations (Heer & Card, 2004). It uses the connection approach also known as a graphing technique that enable the users to manage large hierarchical dataset structures with limited screen display. The algorithms within DOITree calculates to determine which parts of the trees are the most relevant, this helps the readers to identify and navigate through complex hierarchical structures while still maintaining the grasp for the overall dataset visualisation.

Treemap:

First introduced by Ben Shneiderman at the University of Maryland's Human – Computer Interaction Lab (Fowler & Stanwick, 2004), the Treemap is a data visualisation method that enables the users and, or readers to better navigate through the hierarchical structures using nested rectangles that represent categories and subcategories (Johnson & Shneiderman, 1991). The use of colour within the nested rectangles in Treemap is sensitive as it impacts on how the readers perceive the dataset.

Choropleth map:

The choropleth map is described as a visualisation method that utilizes the colour's saturation intensity to display the values in each geographical location, the darker is the colour, the more values and, or data points lie within that particular area. The coloured regions are often cities, countries, or states, and they illustrate the aggregate value, such as the population density, economic growth rate, etc. In addition, Brewer suggested that choropleth map is an essential tool to identify and explore patterns, relationships between variables, and to facilitate comparisons and analysis among different regions (Brewer, 2016).

Analysis

The report will use two datasets for connection approach and enclosure approach, one of which is from the resource available namely WSU courses.xml. The other dataset stored information about medallists at every Summer Olympics from 1896 to 2008, it was chosen from Tableau's public sample dataset collection. Note that both dataset is attached within the report for reference.

1. *WSU courses dataset*

For the WSU courses, due to the nature of the file, there was no need for further cleaning or refining, the file was ready to be manipulated by TabuVis (Figure 1). Via DOITree environment, the collapsable Tree layout was chosen for its distinctive hierarchy nature, enabling the clear view of the parent and child nodes. Interaction was also available, allowing the collapse or expansion of the nodes for further exploration.

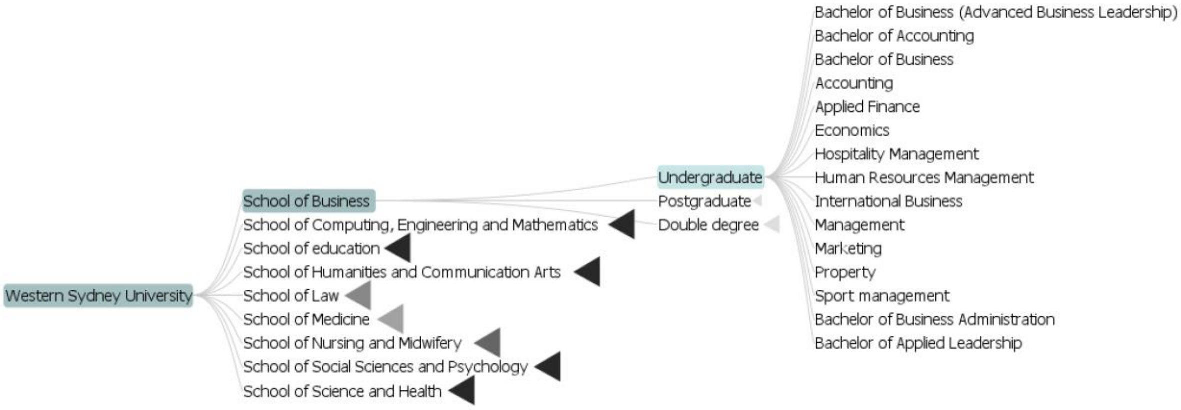
Western Sydney University
School of Business
School of Computing, Engineering and Mathematics
School of education
School of Humanities and Communication Arts
School of Law
School of Medicine
School of Nursing and Midwifery
School of Social Sciences and Psychology
School of Science and Health
Undergraduate
Postgraduate
Double degree
Bachelor of Business (Advanced Business Leadership)
Bachelor of Accounting
Bachelor of Business
Accounting
Applied Finance
Economics
Hospitality Management
Human Resources Management
International Business
Management
Marketing
Property
Sport management
Bachelor of Business Administration
Bachelor of Applied Leadership

Figure 1. DOITree visualisation of WSU courses

The radial tree layout was chosen next to examine the use of display space if the DOITree is to work with large datasets, which can contain more than 100 nodes. It can be observed that there is collision and overlapping of the vertices names and edges, rendering it difficult to make sense of the graph (Figure 2). This leads to the report's next point, using enclosure approach for large datasets.

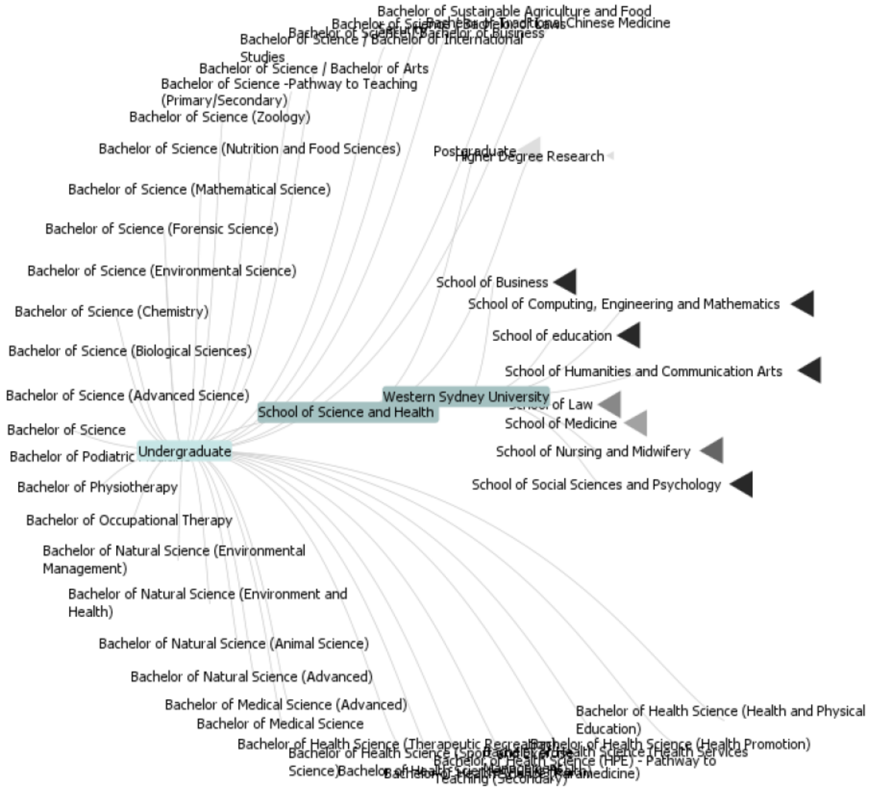
Bachelor of Sustainable Agriculture and Food
Bachelor of Science / Bachelor of International Studies
Bachelor of Science / Bachelor of Arts
Bachelor of Science -Pathway to Teaching (Primary/Secondary)
Bachelor of Science (Zoology)
Bachelor of Science (Nutrition and Food Sciences)
Bachelor of Science (Mathematical Science)
Bachelor of Science (Forensic Science)
Bachelor of Science (Environmental Science)
Bachelor of Science (Chemistry)
Bachelor of Science (Biological Sciences)
Bachelor of Science (Advanced Science)
Bachelor of Science
Bachelor of Podiatric
Undergraduate
Bachelor of Physiotherapy
Bachelor of Occupational Therapy
Bachelor of Natural Science (Environmental Management)
Bachelor of Natural Science (Environment and Health)
Bachelor of Natural Science (Animal Science)
Bachelor of Natural Science (Advanced)
Bachelor of Medical Science (Advanced)
Bachelor of Medical Science
Bachelor of Health Science (Health and Physical Education)
Bachelor of Health Science (Health Promotion)
Bachelor of Health Science (HPE) - Pathway to Teaching (Secondary)
Postgraduate
Higher Degree Research
School of Science and Health
Western Sydney University
School of Business
School of Computing, Engineering and Mathematics
School of education
School of Humanities and Communication Arts
School of Law
School of Medicine
School of Nursing and Midwifery
School of Social Sciences and Psychology

Figure 2. Radial Tree layout of WSU courses

2. *Summer Olympics Medalists 1896 to 2008 dataset*

The dataset was downloaded and stored as a csv file, and data cleaning had to be implemented with Excel, including adding a new column determining the continents to which the countries belong, removing some irrelevant information, such as Excel sheets that did not serve any purposes and several outlying territories on the map. The Treemap was chosen as the layout for primary visualisation on Tableau (Figure 3).

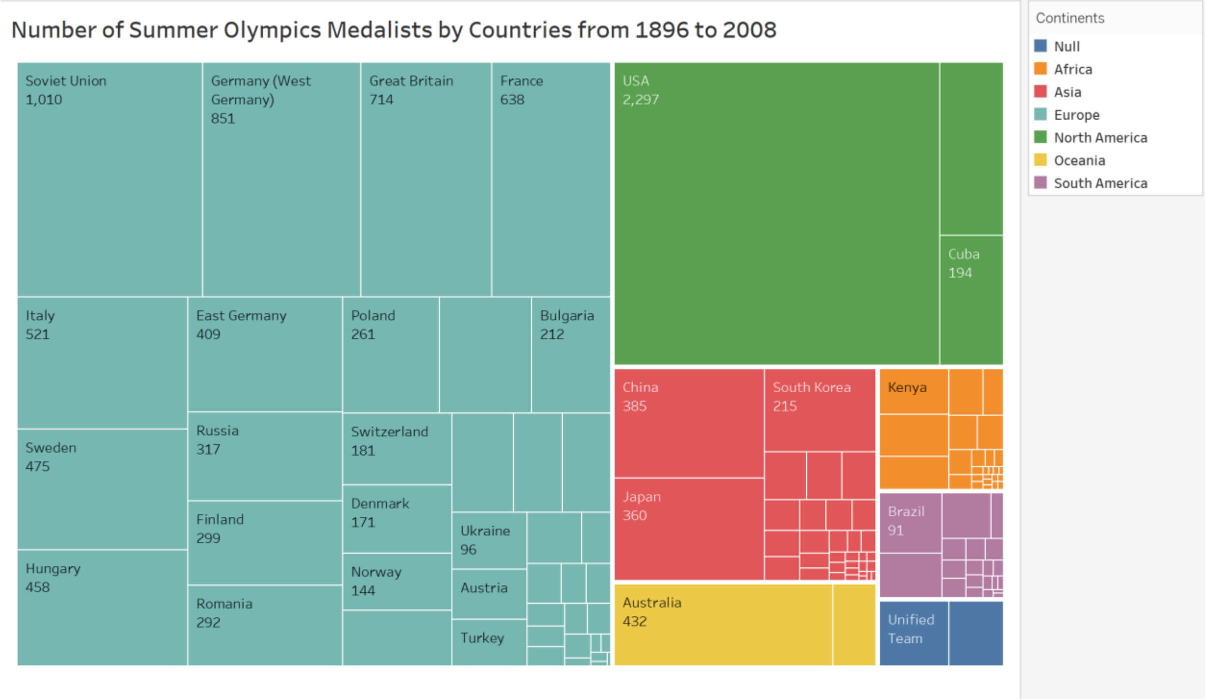

Figure 3. Treemap layout for Summer Olympics Medalists from 1896 to 2008

Finally, the Choropleth map was selected to indicate the number of medals won by countries. This method particularly works well with the dataset for its nature of containing geographical factor and quantitative values (number of medals), which facilitates the demonstration of how the variable changes among the selected areas with colour saturation (Figure 4).

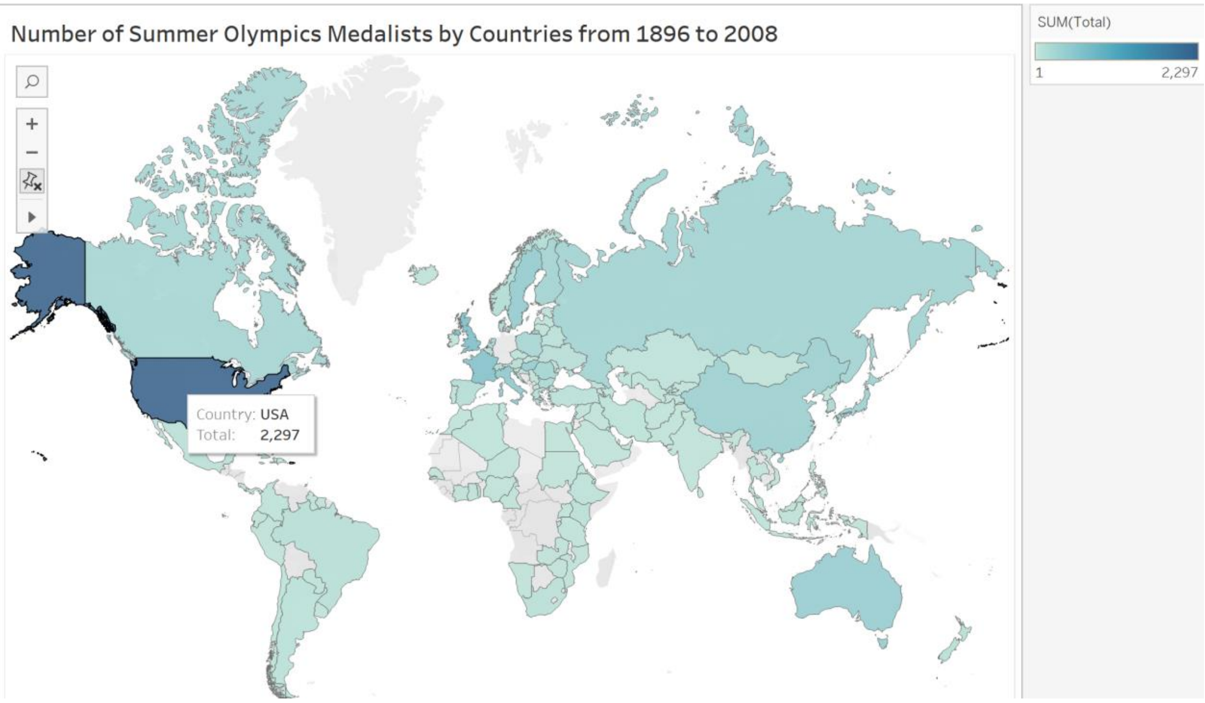
Number of Summer Olympics Medalists by Countries from 1896 to 2008
SUM(Total)
1
2,297
Country: USA
Total: 2,297

Figure 4. Choropleth map layout for Summer Olympics Medalists from 1896 to 2008

Discussion

The branch-based style in the DOITree layout facilitates the overall understanding of the structure with schools as parent nodes and majors as the child nodes. For instance, each school has its own undergraduate and postgraduate, or higher degree research programs, and within each program contained various majors. Furthermore, by assigning weight, or size to the visual cues of each parent node, it is easier to observe which majors have more subsidiaries than the others. However, the inefficiency in display space use is one of the method's disadvantages; research papers from Zhang et al. (2019) and Wang et al. (2020) agreed on the difficulty in analysing large and complex datasets with DOITree because of the displace space's proportional adjustment with the dataset's size and depth. For example, the space used in WSU courses dataset increases by a substantial proportion when expanding parent nodes with more than 20 links (School of Humanities and Communication Arts). The similar pattern is recognised with the Radial Tree layout.

A research paper by Bruls, Huizing, and van Wijk (2000) described Treemap to offer outstanding overview of the extensive hierarchical structures with nested rectangles, while Brewer and Pickle (2002) suggested the effective use of Choropleth map with visualising large datasets. The common feature observed in both graphing techniques is the capability to illustrate aggregated data or encode various dimensions with colour gradient. As an illustration, the number of medallists is the attribute subjected to visualisation, and the colour saturation intensity helps demonstrate which countries have the most or the least Olympics medallists. The pre-attentive nature of the methods also allows readers to find that USA is the country with the most medals won in the given time frame, followed by the old Soviet Union and the European countries.

Conclusion

The report has taken into consideration three different visualisation methods by analysing two datasets with distinctive nature. Each method has its own advantages and disadvantages, their use depends greatly on types of data we are working with. The findings that were discussed in this report may have minimal contribution any particular area; however, the expected goal is illustrating the story behind rows and columns of data, to explore the relationships, and to get a good grasp of the whole picture.

Reference list:

1. Brewer, C. A 2016, *Designing better maps: A guide for GIS users*, Esri Press. https://www.esri.com/en-us/esri-press/browse/designing-better-maps-a-guide-for-gis-users

2. Brewer, C. A & Pickle, L 2002, *Evaluation of Methods for Classifying Epidemiological Data on Choropleth Maps in Series*, Annals of the Association of American Geographers, 92(4), 662–681.

3. Bruls, M, Huizing, K & van Wijk, J. J 2000, *Squarified Treemaps*, In Proceedings of the Joint Eurographics and IEEE TCVG Symposium on Visualization (pp. 33–42).

4. Fews, S 2009, *Information dashboard design: The effective visual communication of data,* Manhattan, NY: Taylor and Francis. https://dl.acm.org/doi/10.5555/1206491

5. Heer, J & Card, S. K 2004, *DOITrees Revisited: Scalable, Space-Constrained Visualization of Hierarchical Data*, ACM. http://vis.stanford.edu/files/2004-DOITree-AVI.pdf

6. Johnson, B & Shneiderman, B 1991, *Tree-maps: a space-filling approach to the visualization of hierarchical information structures, Proceeding Visualization '91*, San Diego, CA, USA, 1991, pp. 284-291, doi: 10.1109/VISUAL.1991.175815.

7. Munzner, T 2014, *Visualization analysis and design*, CRC Press. https://www.cs.ubc.ca/~tmm/vadbook/

8. Nguyen, Q.V & Huang, M.L 2005, *EncCon: an approach to constructing interactive visualization of large hierarchical data*, Information Visualization, *4*(1), pp.1-21.

9. Wang, X, Li, W & Zhang, X 2020, *Visualization of Big Data: From Scalability to Readability*, IEEE Transactions on Big Data, 6(1), 226–239.

10. Zhang, Q, Segall, R. S, & Cao, M (Eds.) 2011, *Visual Analytics and Interactive Technologies: Data, Text and Web Mining Applications*, IGI Global. https://doi-org.ezproxy.uws.edu.au/10.4018/978-1-60960-102-7

11. Zhang, Y, Zhang, J, Li, Z & Wang, J 2019, *An Interactive Visual Analysis System for Hierarchical Data Based on the Degree of Interest Tree*, IEEE Access, 7, 58055–58065.