SCHOOL OF COMPUTING, ENGINEERING AND INFORMATION SCIENCES

Masters Programme in Computing – CG0174

Year: 2010/11

Student Name (ID): Joseph Hogg / 05006060

Dissertation title:

Applying Data Visualisation Techniques to Data Sets Derived from OLAP Cubes Used to Store Data from the National Student Survey

Supervisor: Joe Faith

Second marker: Martin Wonders

Declaration

I declare the following:

1. that the material contained in this dissertation is the end result of my own work and that due acknowledgement has been given in the bibliography and references to ALL sources be they printed, electronic or personal.

2. the Word Count of this Dissertation is 13,288.

3. that unless this dissertation has been confirmed as confidential, I agree to an entire electronic copy or sections of the dissertation to being placed on Blackboard, if deemed appropriate, to allow future students the opportunity to see examples of past dissertations. I understand that if displayed on Blackboard it would be made available for no longer than five years and that students would be able to print off copies or download. The authorship would remain anonymous.

4. I agree to my dissertation being submitted to a plagiarism detection service, where it will be stored in a database and compared against work submitted from this or any other School or from other institutions using the service. In the event of the service detecting a high degree of similarity between content within the service this will be reported back to my supervisor and second marker, who may decide to undertake further investigation that may ultimately lead to disciplinary actions, should instances of plagiarism be detected.

5. I have read the UNN/CEIS Policy Statement on Ethics in Research and Consultancy and I confirm that ethical issues have been considered, evaluated and appropriately addressed in this research.

Signature:

Date:

Acknowledgements

I would like to thank Joe Faith for his advice and support throughout the course of the project and also for allowing me to use the TPP tool. I would also like to thank the following people who, at various points of the project have provided input, allowed me access to resources and have generally made this project possible, they are; Jackie Njoroge, Sally Iles and Rebecca Strachan and also Northumbria University for allowing me to carry out this research.

Abstract

In an age where statistics can tell an organisation more and more about their performance, large organisations are looking for ways to simplify the statistics and present them in a way that is meaningful and useful to them. For a University this is no different. These statistics that can tell people "stuff" are often held in large, sometimes complex data storage structures known as OLAP cubes. For the University studying data from the National Student Survey, knowing how to get this data out of the OLAP cubes and then using it effectively can help to inform which factors affect their students overall satisfaction and something which for a University can be a huge-performance measure and something worthy of further investigation. In this report several techniques are analysed and evaluated with a novel technique known as Targeted Projection Pursuit being chosen as the most suitable method for visualising the data and an OLAP cube generator known as Mondrian used to store the data efficiently. The data has been visualised using the TPP tool and it was found that overall, providing sufficient support and feedback is what students consider the most important factors and crucially what will help gain a "better" score in the National Student Survey.

Keywords: Data visualisation, OLAP Visualisation, Targeted Projection Pursuit, Multidimensional data, National Student Survey

Chapter 1 – Introduction

1. Introduction

Visualising multidimensional data is often a highly desirable task to perform and there are many techniques available for visualising such data. One of these techniques is Targeted Projection Pursuit (TPP) (Faith, 2007). Targeted Projection Pursuit is a general purpose tool that allows for efficient and accurate visualisation of multidimensional data and was the method chosen for this research project to visualise sets of multidimensional data.

Furthermore, one such example of a multidimensional data set that Targeted Projection Pursuit can be applied to is the results of the National Student Survey (NSS). Currently the results of the National Student Survey are held by the client in multiple spreadsheets and comparing the data is a reasonably time consuming and awkward if not sometimes difficult task. Once the data has been collated, the task of actually drawing conclusions from large data sets of numbers and a few limited graphs and pie charts can be difficult.

1.1 Motivation

Given the problems and issues with the current methods used by the client, it was thought that it would have been desirable to solve the following issues; the efficient storage of the large data sets and to provide a way to visualise the data to make it more meaningful to end-users. This report aims to address these issues by proposing a way to store the data in an OLAP cube which can then be extracted and converted to a format that TPP can use, and then create a visualisation of this multidimensional data that is extracted from the OLAP cube.

It was also thought that by using TPP, a better understanding into the factors which most affect the score received for question 22 of the National Student Survey could also be found. The findings of the experiment have been documented in chapter 9.

Further to these reasons discussed above, it is also quite apparent that there is little research work currently being conducted in to this field

(Techapichetvanich and Datta, 2005) and so, it seems that perhaps researching this field further may be worthwhile and contributing to this field of research is particularly appealing.

1.2 Objectives

The objectives of this research are to:

- Research, evaluate and identify suitable OLAP storage solutions (Chapter 4)
- Research, evaluate and identify suitable algorithms and techniques for visualising multidimensional data (Chapter 5)
- Design a solution to solve the research problem (Chapter 7)
- Implement the solution
- Evaluate and test the solution with domain users (Chapter 9 & 10)
- Document and discuss the findings of the research

1.3 Work done

A suitable tool known as Mondrian was identified for creating the OLAP cube. Additionally, Targeted Projection Pursuit (TPP) was chosen as the visualisation tool. The closely-related and common underlying architecture of both Mondrian and TPP (both are built upon the WEKA toolkit) was identified early and it was thought that they would be the most appropriate tools available to conduct the research experiment.

The results of the experiment have been analysed to see whether or not there is any improvement in user-understanding of the data after it has been visualised using the proposed solution. This is measured by testing and evaluating the tool with domain experts who analyse and use the National Student Survey data. There is also a presentation of the results of our hypothesis testing in which it is found which factors are most influential in affecting overall student satisfaction for question 22 of the National Student Survey.

In the report we have researched and evaluated existing OLAP storage solutions, algorithms and techniques for visualising multidimensional data,

designed an experiment to test our hypothesis which we then evaluated and tested and we share and discuss the results and findings of the research.

From the feedback, further conclusions and recommendations for future work were derived and suggested to provide further improvements to the proposed solution for future research work.

1.4 Conclusions

The report concludes that best choices of techniques, research methods and hypothesis have been made and presents an answer to our original research question.

1.5 Structure of the Report

The report is structured in the following way; Chapter 2 introduces and describes the National Student Survey, along with the aims and objectives of the Survey. It also describes the current situation and how the results of the survey are analysed by the Client and addresses what users want to know from the data.

Chapters 4, 5 and 6 provide a survey of the literature. Chapter 4 specifically looks at OLAP visualisation whilst chapter 5 looks at multidimensional data reduction techniques and both identify the tools and techniques that are then used to help to provide a solution to the research problem. Chapter 6 looks at the available tools for building and generating OLAP cubes and gives reasons for the choices that have been made.

In chapter 7, the design of the experiment is discussed and detailed. This chapter also provides a hypothesis, product requirements and a design of the evaluation process.

Chapter 8 follows on from chapter 7 by proposing some design improvements for the TPP tool to make it more usable for end-users.

In chapter 9 there is some hypothesis testing and evaluation of the results of the experiment along with some analysis of the results gathered from the experimental research.

In chapter 10, there is an evaluation of the solution and the overall process that was undertaken during the research project to help inform whether or not the right decisions were made throughout the project.

Finally in chapter 11 there are a series of conclusions and a series of recommendations are proposed to help aid with any future work or research that could be carried out.

Chapter 2 – Visualisation of National Student Survey data

2. Introduction

This chapter provides an introduction to the National Student Survey and how the client currently uses the data from the survey. Section 2.1 introduces the National Student Survey and what its aims and objectives are as well as describing the structure of the survey. Furthermore, section 2.2 of this chapter examines how the stakeholders of this project currently use the data. Finally in section 2.3 there is an explanation of what it is that people want to know and get out of the data.

2.1 The National Student Survey

The National Student Survey (NSS) is, according to its organisers and supporters – IPSOS MORI, the National Union of Students (NUS) and the Higher Education Funding Council for England (HECFE) – a national initiative which asks final year students studying on an undergraduate degree programme, studying at a Higher Education Institution (HEI) in England, Wales and Northern Ireland, as well as certain Higher Education Institutions in Scotland, to give feedback on their course in a nationally recognised format (National Student Survey, 2010). The Survey is based upon the Australian survey known as the Course Experience Questionnaire (CEQ) (Ramsden, 1991) and is almost 100% online, although users may answer the Survey over the telephone or by completing the Survey by hand.

The Survey results can be broken down into three levels, known as JACS (Joint Academics Coding Structure) subject levels. These levels are labelled as Level 1, 2 and 3 where level 3 most closely represents the programmes or departments found in most UK universities.

The survey consists of a standard set of 22 questions (see Appendix D) all of which employ the Likert Scale philosophy, i.e. they ask respondents to answer from a defined set of responses which include: “definitely agree,” “mostly agree,” “mostly disagree,” “definitely disagree” and “not applicable” (Likert,

1932). By asking these 22 questions, the survey organisers aim to assess seven aspects of a students learning experience. These areas are:

- The teaching on a students course
- Assessment and feedback
- Academic support
- Organisation and management
- Learning resources
- Personal development
- Overall satisfaction

(National Student Survey, 2010)

From the responses given by students, each institution is then given a score for each of the seven aspects which have been listed above. Typically, the results which are gathered from the survey form part of the assessment criteria that many of the independent organisations, newspapers and others involved in the production of university league tables use to rank a Higher Education Institution (Times Online, 2008; The Complete University Guide, no date; The Guardian, 2010).

2.2 The current situation

Currently, Northumbria University (sometimes referred to as the University or the Client or the Institute) – one of the project stakeholders – visualise the data in a fairly simple and traditional method using a mix of graphs, pie charts and other pictorial aids, as well as large amounts of typed-data that is simply colour-coded (Njoroge, 2010; Iles 2010) or highlighted, along with some typed data that as been added to a pivot table to allow for easier filtering of the data (Njoroge, 2010; Strachan, 2010; Iles, 2010). This data is collated from various sources including the National Student Survey and its organisers IPSOS MORI as well as a web-based data portal known as Unistats (Njoroge, 2010; Iles 2010).

Through the University's Corporate Planning Department (CPD), the university also has access to a national database system known as the Higher Education

Information Database for Institutions (HEIDI) (HEIDI, 2010). This data, along with the National Student Survey data available to the University, is held in Microsoft Excel spreadsheets and it forms the foundation of the data that is used to generate the graphs and pie charts which the Corporate Planning Department use to monitor performance, make predictions and suggest future trends (Northumbria University, 2010; Njoroge, 2010).

Additionally, the University also uses the HEIDI database systems' in-built functions to produce reports about the Institute. These reports are known as Taylor Squares (HEIDI, 2010). Taylor Squares (See figure 1) are named for Brian Taylor, the inventor of the technique and have been used in HEIDI since 2008 (Phelps, 2010).

Figure 1 shows a Taylor Square. This particular example compares the student load as a percentage and by each of the subject areas that students answer questions for in the National Student Survey. Furthermore, it has been filtered even further to compare the University only against other universities that are part of the University Alliance (a group of 23 universities of which Northumbria University is a member) (University Alliance, 2010).

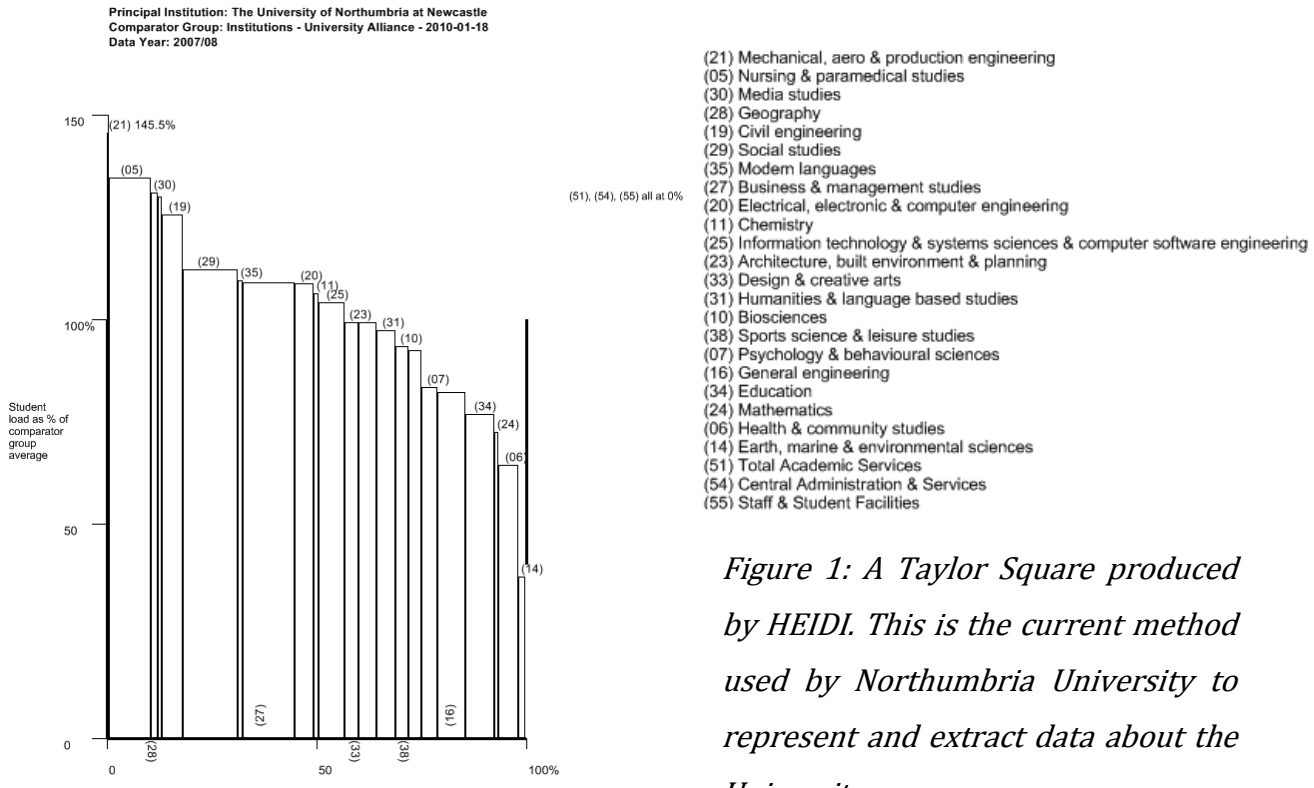


Figure 1: A Taylor Square produced by HEIDI. This is the current method used by Northumbria University to represent and extract data about the University.

The university collates data from HEIDI, the NSS and other sources and then packages it all together into something known as a data cube or OLAP cube for later exploratory analysis. The output from the data cube can then aid the CPD in making judgements (amongst other things) about whether or not targets have been met or not, and if progress is being made. The data can also sometimes be used to monitor, evaluate and compare the university's performance against other Higher Education Institutions within the United Kingdom.

Given the widespread and sometimes unconnected sets of data which are used to monitor performance, the task of comparing the university against another institution can be challenging using the current methods.

It is this problem of unconnectedness and widespread nature of the data that is looking to be solved during this project. The University for its part, wishes to employ a new method which makes the task of presenting the data easier.

2.3 What people want to know from the data?

There are several things that the university wishes to learn from undertaking such an exercise. As has been briefly mentioned in Section 1.2, one of the main reasons for collecting this data and analysing it is to gain an overview of where the university ranks compared to other HEI's.

Like many activities, the pursuit of gaining further knowledge is a major reason for wanting to analyse and understand the data that all of the sources (HEIDI, NSS scores etc.) provide. This pursuit of further knowledge can then in turn lead to adding value to a particular process or operation.

The main points or "things" that people want to know from this data can be summarised in the following few points;

- What factors affect how well an institution performs?
- What are "better" (i.e. higher-ranked) institutions doing that they aren't?

- Why are other institutions better at something than they (the Client) are?
- What can be done to improve their own ranking and to meet their targets?
- What would be the affect on the University and its ranking if they were to do something differently? And;
- What are they doing well that other HEI's aren't doing so well?

Furthermore, the University would like to be able to conduct horizon scanning to see what competitor universities are doing and how they are doing compared to the University, but also looking at what so-called "aspirational universities" are doing – these are universities that the client aspires to be like.

2.4 Research decisions

After careful consideration and discussion with the stakeholders it has been decided to narrow the scope of the research problem. Rather than try and collate all the data that is available to the university such as financial data, NSS data and data from HEIDI, only the NSS data will be visualised and focus at this stage will be emphasised on what factors from that particular data set most affects overall satisfaction. This is discussed and made clearer later in the report in section 6.1 (Hypothesis).

2.5 Summary

This chapter has introduced the National Student Survey. The aims, objectives, purpose and structure of the survey have been described and the project stakeholders. The reasons for examining the data and what people want to know from it have also been discussed. It also has defined and narrowed the scope of the research experiment. In the next chapter the method of data visualisation is reviewed and disscussed.

Chapter 3 – Data visualisation

3. Introduction

In this chapter the general idea of data visualisation is introduced and discussed and several examples are provided to illustrate the types of graphical representations that can be realised using data visualisation techniques.

3.1 Data visualisation

According to Friendly & Denis (2001), Data visualisation is, "the science of visual representation of data." Thus, data visualisation is when data has been extracted and represented in some schematic form and displayed by representing all of a data's attributes and variables. The University of Alberta (1999) defines data visualisation as the "visual interpretation of complex relationships in multidimensional data." It is often sometimes seen as the process of transforming statistics into knowledge (Office for National Statistics, no date).

Its origins are in the early day of computer graphics, in the 1950's, but it wasn't until the National Science Foundation (NSF) published a report entitled *Visualization in Scientific Computing* (NSF, 1987) and the immense and rapid growth in size of data sets occurred, that the need for a new technique was stressed. The report by the NSF spurred on the creation of a number of research activities and the birth of data visualisation as a specialism within the field of computer graphics (Post et al, 2003).

The main purpose of data visualisation is to communicate information in a clearer and much more effective way by using graphical representations and as such is a form of visual modelling (Freidman, 2008). It is often closely related to the disciplines of data acquisition, data analysis, data management and data mining. The limits of the applications of data visualisation are really unbounded it would seem and the growth of the technique seems to be growing in popularity.

The main applications of data visualisation can be found in many industries, but particularly in business and finance, statistics, administration and digital media (Post et al, 2003).

The Bloom Diagram is an example of a data visualisation tool and allows users to visualise the contributions of particular individuals to open source projects (Cheng & Kerr; IBM, no date).

Another notable example of data visualisation in action is *Newsmap*, which sizes news stories based on their popularity and clusters similar stories to make patterns more recognisable (Weskamp, 2004).

TIME magazine also have produced a diagram that uses data visualisation techniques. In it they produced a visualisation that used the notion of spikes to denote areas of high population densities (TIME, inc, 2006).

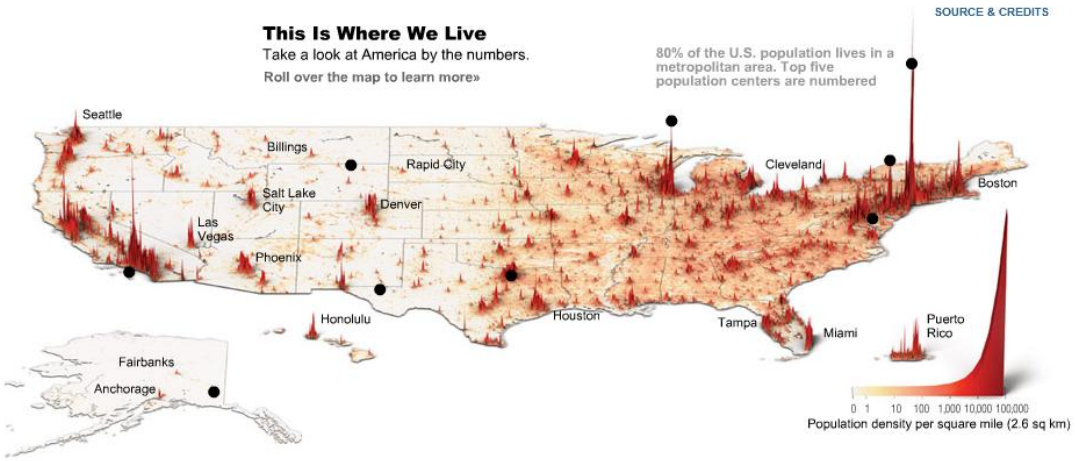


Figure 2: An image taken from TIME magazine showing the population density of the United States of America. Areas with large population densities have taller spikes, such as Boston and Miami. (Time, inc, 2006)

Of course there are numerous other examples available as data visualisation is a large-sprawling field and has many applications and the previous examples are just shown as examples of data visualisation.

3.2 Summary

This chapter introduced the idea and philosophy of data visualisation and provided several examples of data visualisation in everyday use such as Newsmap that resizes new stories based on popularity.

Chapter 4 – OLAP Visualisation

4. OLAP Visualisation

In this chapter OLAP visualisation is discussed and the reasons why we may wish to visualise such data is presented in section 4.1. The chapter also describes the current work into OLAP visualisation.

4.1 Why visualise OLAP data?

With the general growth of statistical data being made available in databases, it is not surprising that there are efforts to capitalise on as much of this data as possible and in doing so, gain a better-understanding, knowledge and understanding of the importance of the data.

While it is relatively simple to visualise small sets of data, for example, two-dimensions; costs versus growth or sales versus profits, even three dimensions such as sales versus growth versus profits can be visualised quite sufficiently, the problem arises when the data is multi-dimensional, i.e. more the 3 dimensions. Visualising these sets of data becomes much more of an issue and, as such, the methods we may use to represent 2 or 3D spaces graphically – such as pie charts and bar graphs – are not particularly useful and rendered useless in some cases (Wilkinson, 2005, p. 248).

OLAP visualisation can help solve this problem. It can help solve this problem because; 1. It is multi-dimensional in form meaning we can visualise many features or attributes of data, in other words more than three-dimensions and 2. It also allows us to show a snapshot of data over time which is often very desirable when considering and make judgements using information gathered from statistical data as often people want to know where they have come from and where they are going to.

Figure 3 shows a basic example structure of an OLAP cube. It has 3 dimensions (the multidimensional element) which are “Products,” “Cities,” and “Time” (which also satisfies the time element)

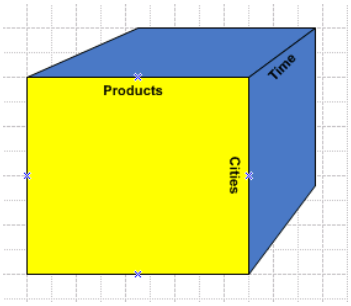
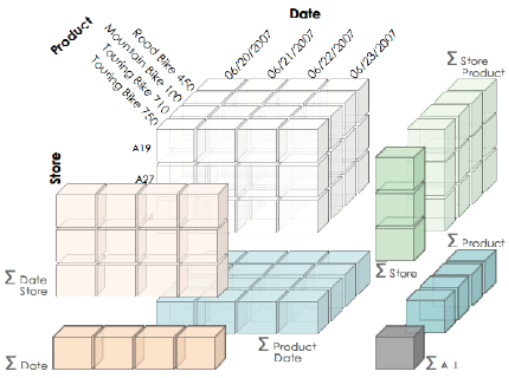


Figure 3: A simple representation of an OLAP Cube with 3 Dimensions

(Roeder, 2008)

4.2 Current work into visualising OLAP data

In their paper, Cuzzocrea and Mansmann (2009), acknowledged that to date there has little work currently done in the field of OLAP visualisation but state that visualising multi-dimensional data is attracting the attention of a wide and multi-disciplinary community of researchers and practitioners (Cuzzocrea et al, 2009). In their paper they summarise the principle models currently in use to visualise OLAP data, citing the CoDecide Model (Gebhardt et al, 1997), the CPM Model (Maniatis et al, 2003), Tableau (Hanrahan et al 2007), DIVE-ON (Ammoura et al, 2001) and the Hierarchical Dynamic Dimensional Visualization (Techapicetvanich & Datta, 2005).



		Measures							
		Quantity				Amount			
Dimensions		Store							
Product	Date	A19	A27	A34	Total Store	A19	A27	A34	Total Store
− Road Bike 450	06/20/2007	2	7	4	13	498	1743	996	3237
	06/21/2007	9	12	10	31	2241	2988	2490	7719
	06/22/2007	3		7	10	747		1743	2490
	06/23/2007	5	1	9	15	1245	249	2241	**3735**
Total Road Bike 450		19	20	30	69	4731	4980	7470	17181
− Mountain Bike 100	06/20/2007	8	10	3	21	6392	7990	2397	16779
	06/21/2007	5	11	4	20	3995	8789	3196	15980
	06/22/2007	9	7		16	7191	5593		12784
	06/23/2007	6		4	10	4794		3196	**7990**
Total Mountain Bike 100		28	28	11	67	22372	22372	8789	53533
− Touring Bike 710	06/20/2007	5		9	14	2995		5391	8386
	06/21/2007	7	2	12	21	4193	1198	7188	12579
	06/22/2007	4		13	17	2396		7787	10183
	06/23/2007		2	8	10		1198	4792	**5990**
Total Touring Bike 710		16	4	42	62	9584	2396	25158	37138
+ Total Touring Bike 750		19	12	15	46	10621	6708	8385	25714
Total Product		82	64	98	244	47308	83765	49802	133566

Figure 4: A three-dimensional data cube shown left and the accompanying pivot table

(Roeder, 2008)

Work on the Tape Model, suggested by Gebhardt et al (1997), was one of the earliest attempts to develop a tool to visualise high-dimensional data and builds on the work that Codd et al (1992) carried out. They present their tool (CoDecide) that uses the metaphors of tapes and tracks. Whilst their tool was relatively simple in design and not much more that an advanced Microsoft Excel pivot-table, it was a step up from previous models and began to truly tackle the issue of OLAP visualisation.

Maniatis et al (2003) took Gebhardt et al's (1997) idea, building upon the idea of tapes, but took a somewhat different approach with their model, proposing a tool that separates the data into two layers; logical and presentation. Their tool, the *Cube Presentation Model* (CPM), they argued, was better adapted for application to advanced visualisation of OLAP data, stating that the separation of the presentation of data and retrieval of data is more efficient and maps much closer to the reality of how users work with data; something they describe as working in "sessions of queries rather than sequences of unrelated queries," (Maniatis et al, 2003).

Some argue, that unlike most of the attempts to visualise OLAP data, users should be allowed to combine several different visualisation techniques to get the best spread (Hanrahan et al, 2007) rather than the normal methods which most models use, such as providing pre-defined templates, visual formats, widgets and wizards (Cuzzocrea et al, 2009). Hanrahan et al (2007) say that their implementation, known as *Tableau*, offers "high expressiveness" by using the "most popular" and "proven" visualisation techniques and combining these to provide many visual aids to users to help gain a better understanding. However, the suggestion by Hanrahan et al is argued by several researchers including Tegarden (1999) who say that it is better develop new, more novel and less well-known techniques to take advantage of the multi-dimensional and hierarchical properties of the data.

Work by Ammoura et al (2001) on their DIVE-ON project offers, claim Cuzzocrea et al (2009), "one of the most unique experiences in the OLAP visualization research field," as it exploits our natural human ability to interact with spatial objects. In exploiting this natural behaviour – using colour, shapes and size to represent and portray ideas and information – when visualising information, "the knowledge fruition" phase is more productive and "intelligently" exploits the OLAP technology layer to efficiently support the multidimensionality of data.

The other major area, most current method of OLAP visualisation is interactive visualisation. The Hierarchical Dynamic Dimensional Visualization (HDDV) (Techapicetvanich & Datta, 2005) model is an example of this kind of research. In their model, as the names suggests, Techapicetvanich and Datta introduce the idea of a more hierarchical structure that has dimension hierarchies which are grouped into bars based on their measurement value for the particular dimension they are part of. Techapicetvanich and Datta, unlike in a similar project first proposed by Sifer (2003) do not explicitly link the bars and in doing so can preserve the divided data more effectively.

4.4 Summary

From the literature, it is clear that there is little research into the area of OLAP visualisation at present. Disagreement between researchers and research groups also means that a clear way forward is yet to be found. Whilst most argue that new and more novel methods need to be found, some believe that those wishing to visualise OLAP data should use a combination of tested and proven methods to produce a more accurate and better reflection of the data.

From these findings in the literature survey it seems clear that there is scope for further research into this area that addresses the issues of; the best forms of graphical representation to use when visualising OLAP data and how the OLAP data can be extracted and then presented, but in particular the way in which

data over time is handled is an area that will make for an interesting investigation.

Chapter 5 – Dimension reduction for multidimensional data

5. Introduction

This chapter examines the different algorithms that can be employed to reduce the dimensionality of multidimensional data. Two main theories that are presented are; linear projection in section 5.2 and non-linear projection in section 5.3. Each technique is examined and several examples of each method are provided including Principal Component Analysis (PCA) (section 5.2.1), Projection Pursuit (section 5.2.2), Targeted Projection Pursuit (section 5.2.3) and Multidimensional Scaling (MDS)(section 5.3.1) .

5.1 Multidimensional data

Multidimensional data is commonly visualised in two ways. It can either be visualised either linearly or nonlinearly. The task of visualising multidimensional data often involves a process known as “projection” on (Wilkinson, 2005, p. 248) and involves projecting the data onto a projection plane.

5.2 Linear projection

The first and most widely used approach is to use linear projection methods. In linear projection, matrices are constructed that can be multiple in their quantity and are referred to as the transformation matrices. In a linear projection, the multidimensional data can have the following operations performed on it; it can be: transformed or translated, rotated and scaled (Lang, 1984).

In linear projection, the principal idea is that a set of matrices are produced. The matrices are then multiplied together to give a value as the output which is then used to place and position an attribute on the projection.

In the multiplication process, the columns of the first matrix are multiplied by the rows of the second matrix and the summation of the two placed in another matrix to give the value that will be used to position the value on the on the projection plane. As the name suggests, linear patterns are also very easily spotted if this projection process is used to project data.

There are many linear projection methods, some of which include projection pursuit (Freidman & Tukey, 1974), principal component analysis (PCA)(Pearson, 1902), as well as a novel technique known as Targeted Projection Pursuit (TPP) (Faith, 2007).

5.2.1 Principal Component Analysis (PCA)

Principal Component Analysis is a technique originally developed by Karl Pearson (1902) to reduce the dimensionality of a data set which contains large numbers of interconnected variables and at the same time maintaining as much of the variation as possible that exists in the data (Jolliffe, 2002). PCA achieves this by transforming the data to a new set of variables known as the principal components (PCs). These PCs are unrelated and ordered in such a way that the first few variables of the set maintain most of the variation that exists in all of the original variables (Jolliffe, 2002). For example a variable set that has had dimension reduction performed on it will form a new set of variables, say { $\boldsymbol{x}_1$, $\boldsymbol{x}_2$, $\boldsymbol{x}_3$... $\boldsymbol{x}_n$} where $\boldsymbol{x}_1$ will contain the most variation and $\boldsymbol{x}_2$ will contain more variation than $\boldsymbol{x}_3$ and so on. The first variable (with the most variance) is referred to as the first principal component.

An example of this can be explained as follows, using a teapot as example. A teapot is a 3-dimensional object and our task is to view the teapot in such a way that we can see the most visual data possible.

Asked to choose the best view of a teapot to gain the best visual in-sight in to the data, most people would choose to view the teapot from above (figure 5.1), however this is not the case.

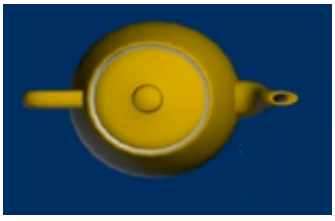


Figure 5.1: A image of a teapot from above. Commonly the view people choose thinking it offers the best view of the teapot as a whole (VisuMap, 2009)

We can actually do is to use PCA to allow us to rotate the teapot around its centre and find the longest axis (figure 5.2).

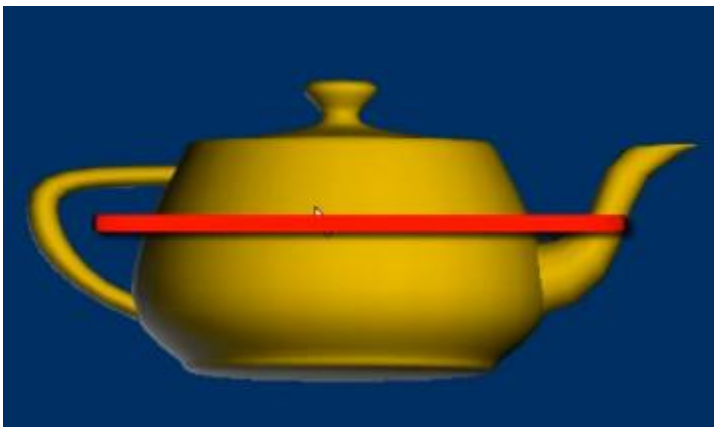


Figure 5.2 (VisuMap, 2009)*: the same teapot as in figure 5.1 shown after the PCA algorithm has rotated the teapot around its centre and found the longest axis (the red line). This red line would be our* x_1 value or principal component in our matrix.

After finding the longest axis – our principal component – we next find the second longest axis or our x_2 value (figure 5.3).

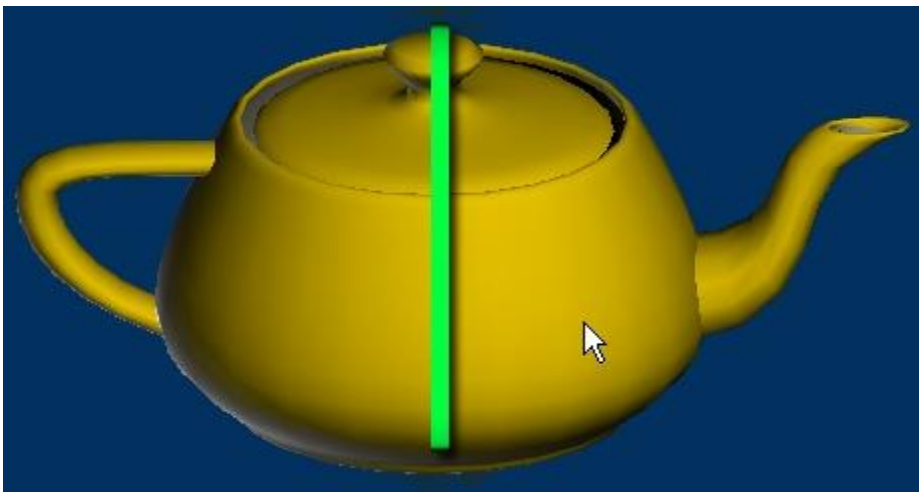


Figure 5.3 (VisuMap, 2009): the same teapot as in figure 5.1 & 5.2 shown after the PCA algorithm has rotated the teapot around its centre and found the second longest axis (the green line). This green line would be our x_2 *value or in our matrix.*

5.2.2 Projection Pursuit

Projection Pursuit (PP) is a technique developed by Friedman and Tukey (1974). Their method aims to seek out "interesting" linear projections and plot them on a plane.

Such is its design, projection pursuit avoids the problem of the "curse of dimensionality," a common issue that many multivariate techniques for visualising multi-dimensional data cannot avoid. This "curse of dimensionality" problem occurs when we have a high-dimensional space that is largely empty. Projection pursuit avoids this by working in low-dimensional linear projections (Huber, 1985).

Furthermore, what projection pursuit also offers is the ability to ignore irrelevant variables that may be noisy or offer poor informational value (Huber, 1985).

5.2.3 Targeted Projection Pursuit

Targeted Projection Pursuit (TPP) (Faith, 2007) is a general purpose tool for visualising high-dimensional data sets. It is similar in ways to projection pursuit and is an example of a linear projection, however as the name suggests it allows for more specific targeting of the data. It is different in that Targeted Project Pursuit hypothesises the most ideal view of the data and then proceeds by finding a projection that best fit that view (Enshaie & Faith, 2010).

TPP offers the advantage that it allows a user to test a hypothesis. A user can play with the view of the data by "grabbing" the data and create different projections of the data. If a projection can be found then the projection is displayed and visual feedback is given to the user, otherwise the data points will remain in-situ if a new projection cannot be found (Faith, 2007). Points can be added to TPP without having to re-map the whole projection

Additionally, what TPP also does that is novel amongst other similar techniques is that it exploits the users own ability to recognise patterns in data (Faith, 2007). This ability that we all possess as humans to recognise patterns means

that much of the work can be conducted with a fair-amount of accuracy while the underlying algorithm that TPP uses, can carry out the other work including the work needed to display the points in the correct position on the projection.

Finally, TPP is built upon the WEKA toolkit, a well-known and widely used data-mining technique. This coupled with the fact that the decision to use the Mondrian cube builder (discussed later) has been chosen means that one can easily import the data in ARFF into TPP.

5.3 Nonlinear projection

Nonlinear projection can be used in several ways to visualise multidimensional data. Wilkinson (2005) describes the first of these as relaxing the global structure of the multidimensional data to reveal interesting parts of the data. Banchoff (1996) describes several methods that involve visualising geometric objects using nonlinear projection.

5.3.1 Multidimensional scaling

The most common approach used in nonlinear projection is known as Multidimensional scaling (MDS). MDS scales data to a very high dimensional level (Faith et al, 2006) and unlike in projection pursuit, it does not "pursuit" projections but more that it is a map-based model (Borg et al, 2005). MDS maps the data on to the projection plane. The proximity of points to one another gives an indication of how similar or "preferred" they are to one another (Cox et al, 2001).

So, Borgatti (1997) states an example as follows: take for example several brands of air fresheners. Using MDS, brands that are perceived to be similar to one another are plotted near one another whilst brands perceived as being very different from one another are plotted far apart from one another.

A simplified view of the algorithm is as follows:

1. Assign points to arbitrary coordinates in p-dimensional space.

2. Compute euclidean distances among all pairs of points, to form the Dhat matrix.

3. Compare the Dhat matrix with the input D matrix by evaluating the stress function. The smaller the value, the greater the correspondance between the two.

4. Adjust coordinates of each point in the direction that best maximally stress.

5. Repeat steps 2 through 4 until stress won't get any lower.

Borgatti (1997)

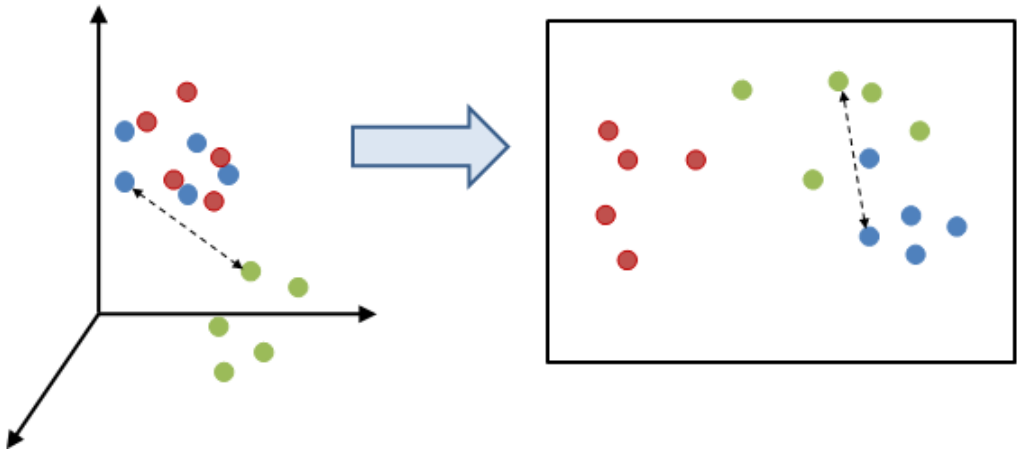


Figure 6 shows an example of data that has been projected using a non-linear method; Multi-dimensional scaling. It shows the data from two viewpoints on the projection. (Faith, 2010).

MDS maps the data and then depending on the proximity of points to one another gives an indication of how similar or "preferred" they are to one another (Cox et al, 2001).

5.4 Linear projection versus non-linear projection

The major differences between linear and non-linear projection can be described as follows; non-linear projection can be inaccurate but is able to show

all relationships between items being projected. Linear is the opposite, it is accurate when only relationships between items are partial.

Additionally, with non-linear projection data must be re-mapped each time a new point is added, in linear projection relationships can be shown as points are added without having to re-map.

Furthermore, non-linear projection cannot handle the "curse of dimensionality" effectively unlike linear techniques such as projection pursuit (Huber, 1985) which can cause problems, and the number of non-linear techniques available to choose from are more limited compared to the number of linear projection methods of which there are many (Faith, 2010).

5.5 Choice of technique

From the review of the techniques available to visualise multidimensional data, it has been decided that using Targeted Projection Pursuit (Faith, 2006) is the most suitable. TPP is closely built on the WEKA toolkit which in turn is closely linked with the chosen OLAP cube builder (Mondrian)(both discussed in the subsequent chapter). The 3 tools can be used together to make the visualisation process easier. Furthermore, TPP has been proven to be the most successful algorithm of its type at visualising multidimensional data when compared with greedy algorithms and filter algorithm techniques (Enshaie & Faith, 2010). TPP provides us with a pre-built tool and algorithm that we can use to visualise the NSS data. Using TPP also means that the scope of the project does not detract from the main purpose which is to visualise and analyse National Student Survey data. If we were to develop a linear or non-linear algorithm and tool from scratch this would be time-consuming and a project in itself. Therefore it is acceptable to use what is already there rather than re-inventing something over again.

5.6 Summary

In this chapter a description of the available techniques that could be used to visualise multidimensional data. Although the techniques have advantages and

disadvantages one suitable method has been identified. The chapter has identified Targeted Projection Pursuit (TPP) as the most suitable and successful method to visualise the data. The choice of method is evaluated in chapter 8. The next chapter discusses the design and structure of the experiment undertaken.

Chapter 6 – OLAP cube builders

6. Introduction

This particular chapter overviews the available OLAP cube builders currently available for use, it also evaluates each tool for fitness for purpose and gives a brief description of each of the tools.

6.1 Pre-built tools for visualising OLAP data

Aside from all of the research into OLAP visualisation, there are many good examples of tools that can build OLAP cubes and then allow the data to be formatted in several ways to allow the data to be visualised in other tools. Some of these tools include Mondrian (Pentaho, 2010).

Mondrian allows users to build OLAP cubes and then perform analysis on the data sets in an easy way. It is built upon the WEKA toolkit (Pentaho, 2010) – a widely-used, open source data mining tool kit (University of Waikato, no date) - and has been developed in Java.

Essentially it is a data analysis tool that provides a way to organise data effectively and create an ARFF file which is compatible with the WEKA toolkit. What it does not do is provide a visual representation method for visualising data.

In addition to Mondrian, there are several other OLAP visualisation tools available. One of these is BI-Lite CUBE-it Zero which has been created by BI-Lite (2010). The tool is similar in functionality to Mondrian and allows for the easy creation of OLAP cubes that once built can then be analysed. However what it does lack is the ability to convert the data into a text format that could potentially be exploited by another visualisation tool, i.e. unlike Mondrian we can not get output in ARFF format.

Finally, there is a tool which many of the other OLAP tools are built upon and that is the WEKA toolkit. Developed by Machine Learning Group at the

University of Waikato (no date). It has been developed in Java and offers more than just visualisation capabilities and does allow users to connect to SQL databases University of Waikato (no date), however it is not particularly built towards creating OLAP cubes, it is more a tool for modelling and data analysis.

Of course, one could create an OLAP generation tool from scratch, however this is seen as beyond the scope of the project and therefore it is seen as acceptable to use a pre-built tool to handle this element of the research experiment.

To aid with choosing an appropriate the following table shows the criteria used to evaluate each of the tools.

Criteria	OLAP cube generator		
	Mondrian	**BI-Lite**	**WEKA**
Ease of use	Easy to use	Easy to use	Easy to use
Compatibility with TPP	Yes	No	Yes
Availability	Open-source	Open-source	Open-source
Ability to import the raw excel data	Yes	Yes	Yes

Table 1: showing the criteria used to evaluate the OLAP cube generators

It is apparent that using a pre-built tool, i.e. Pentaho's Mondrian tool to generate an OLAP cube will be the best forward. Rather than focussing on a method to generate the cube from scratch, using a pre-built tool means that the focus can be shifted to concentrating on ways to visualise the data graphically. The scope of the project does not provide enough time tools to generate OLAP cubes nor does it appear that this aspect of the project would be the most useful or interesting to the client or researcher.

With this in mind the decision has been made to use Pentaho's Mondrian tool for generating our data cube. The reasons for choosing this tool over the others are due to its ease of use, and compatibility with our visualisation tool, i.e. TPP

added to the fact that both the cube builder and TPP are both built on the WEKA toolkit and can both handle ARFF files mean that it seems the most sensible choice of tool. Whilst WEKA and BI-Lite's tool have advantages, their offerings are either too far from what we actually want, i.e. the WEKA toolkit does not exist as an OLAP cube builder as it's main purpose, whilst BI-Lite's tool is not compatible with the visualisation tool.

6.2 Summary

This chapter has presented several possible tools that could be used to generate our OLAP cube. There has been an evaluation of each of the tools and compares each it turn and provides a reasoning for the chosen method.

Chapter 7 – Experimental Design

7. Introduction

In this chapter the experimental method is presented in section 5.2. Section 5.1 presents a hypothesis which is then evaluated in chapter 7. There is also a list of the product requirements in section 5.3

7.1 Hypothesis

Given what has been found during the literature survey the following hypothesis has been formulated:

Targeted Projection Pursuit (TPP) can be used to visualise historical National Student Survey (NSS) data from an OLAP cube and identify which factors most influence students overall satisfaction (Q22 of the National Student Survey) between course programmes and over time.

7.2 Experiment method

The experiment involves taking National Student Survey data and extracting it from an OLAP cube. The OLAP cube holds the National Student Data as numerical values which represent a percentage of the students that answered each of the 22 questions, and also how they responded, i.e. satisfied, not satisfied...etc. After creating the cube, the data can be extracted and converted to a format that is readable by the visualisation tool and a trajectory of the data plotted, i.e. a path of the data over time. Once we have our trajectory, we can then find a view that distinguishes courses and years and discover which factors influence overall satisfaction (Q22) and lead to a high or low score for question 22 of the National Student Survey and compare the results over time.

7.2.1 Data

The data provided by the university is a set of numerical values that represent the responses of the students for each of the 22 questions from the National Student Survey which in turn has been supplied by the organisers of the survey via the University.

The following details are held in an Excel spreadsheet about each of the programmes that students have given feedback on:

There is a subject column and 22 columns 1 for each of the questions in the NSS survey which is a value that represents an average response value for each question. Our Excel spreadsheet is made up of 23 columns (1 subject column which identifies the subject area and 22 columns, each one with a value for the average response for each of the 22 questions in the survey).

For each year of the survey, our Excel file has 19 rows made up of a row for each of the subject areas which map onto the JACS subject areas defined by the Higher Education Authority (HEA). JACS subject areas are the most commonly-aligned subject identifier available and are used in the National Student Survey to map results to subjects. The JACS subject areas are also what the University uses to map results to individual schools. For example, the JACS subject area "(L112) Law" would represent the result set for the University's Law School and all the courses in that particular school.

The data can be broken down in to 3 levels (see section 2.1). The lowest that the JACS results can be broken down to JACS level 3, this is the level at which our JACS level 1, the data would be far too general to make assumptions about or to perform adequate analysis on.

Each of these values mentioned act as our attributes which are used by our visualisation tool to determine the position of the data points (or attributes) on our data projection. The data set which is being included, spans a period of 4 years from 2006 - 2010. A sample of the data set can be found in Appendix E

7.2.2 OLAP cube

The OLAP cube creation is handled by Mondrian an open-source OLAP server that allows for the easy conversion of a Microsoft Excel file to an OLAP cube. Mondrian presents us with a pre-built cube build that can be used to avoid the need to spend time constructing a cube from scratch.

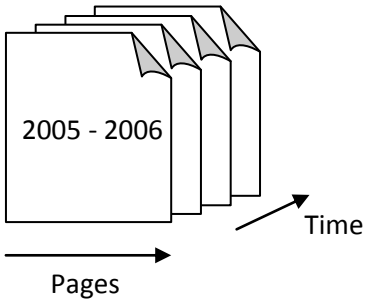
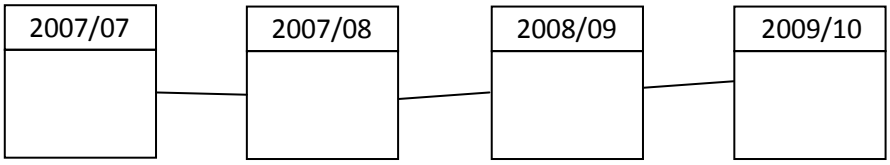


Figure 7: A generalised design of the OLAP cube with pages of data over time

Our cube is structured in the following way and contains four pages or tables of data, each representing 1 year of the National Student Survey results.



The table structure of our OLAP cube

Furthermore, our cube builder is compatible with the WEKA toolkit a data mining software kit (University of Waikato) that allows for exploratory analysis of the data. Mondrian allows for the conversion to an ARFF file which is used by WEKA software and by our visualisation tool; TPP.

Figure 8: The process of creating the cube using Mondrian

7.2.3 Extraction

Using Mondrian allows us to convert the OLAP cube into an ARFF format which can then be exported and the data extracted from the cube.

With our ARFF file we can then present it to our visualisation tool – TPP – which can then read in the ARFF file ready for visualisation. Once the data is extracted into ARFF format we can then connect our cube to our visualisation tool and load it and ready it for visualisation.

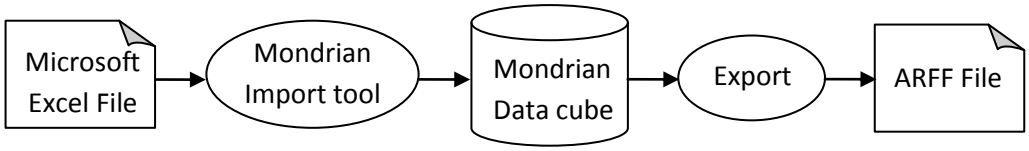


Figure 9.1: The process of creating and extracting/exporting the data cube

7.2.4 Visualisation process

Following on from the construction of our OLAP cube and the extraction of the data into an ARFF file, we can begin visualising our data.

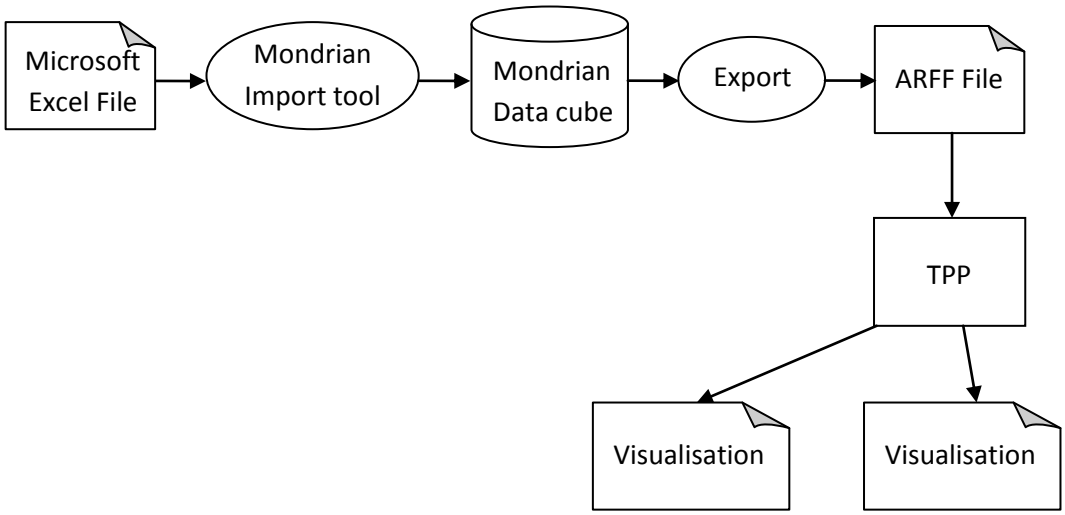


Figure 9.2: The whole process of creating and extracting/exporting the data cube and then visualising using the TPP tool

The software tool we are using (TPP) is built upon WEKA, as is our cube builder, so it makes good design sense to use these two tools. Using TPP we can load in the ARFF and visualise the data. Additionally TPP has been proven to be as efficient as and/or more effective at visualising data than other similar methods (Enshaie & Faith, 2010).

TPP then allows us to colour the data points by particular attributes of the data, in our case we choose to colour the points by responses to question 22 of the National Student Survey and also to separate the points by question 22 of the

survey. This in turn allows us to determine the dominant axes, i.e. the ones that make the most difference or have the greatest significance.

7.2.5 Results evaluation

As part of the research the results will be evaluated to see whether or not the criterion set out at the beginning of the project has been met.

The results will be presented to domain experts who will be asked to evaluate whether the project answers the questions they need answering and whether it provides an improvement on previous methods and techniques currently used to try and carry out this work.

The evaluation will take two forms; there will be some:

- Semi-structured interviews with end-users to get feedback
- User testing

The kind of questions that will be asked will be:

- Does the visualisation improve understanding?
- Is the method more useful than previous methods?
- What elements of the tool do you find particularly useful
- What elements of the tool do you find are difficult to understand or inappropriate

7.3 Product requirements

The following is a list of requirements that needed to be satisfied to complete and implement the solution to our problem. There are few requirements in terms of the software elements of this project as already there is several software components already pre-built. The following list the general requirements of the project:

- A data cube
- A tool to visualise data
- A method to connect TPP to the data cube

7.4 Summary

This chapter has presented the experimental design adopted and discusses the method that will be used to evaluate the results of the experiment. Finally a list of product requirements has been drawn up which will be required to conduct the experiment. The next chapter looks at the design issues and requirements of the TPP tool and solutions for improving the usability of the tool. It will also capture the requirements of the users.

Chapter 8 – Requirements analysis and design

8. Introduction

In this chapter the current design of the visualisation tool (TPP) is presented and evaluated. Design improvements are suggested and potential solutions are evaluated in turn. Section 6.2 provides the user requirements. In Section 6.5 there is a proposed design solution presented.

8.1 Current design overview

The current deign of the TPP tool is a Java GUI and has several design issues which need to be addressed. As the tool has been developed further and enhanced, it has become more cluttered and less useable.

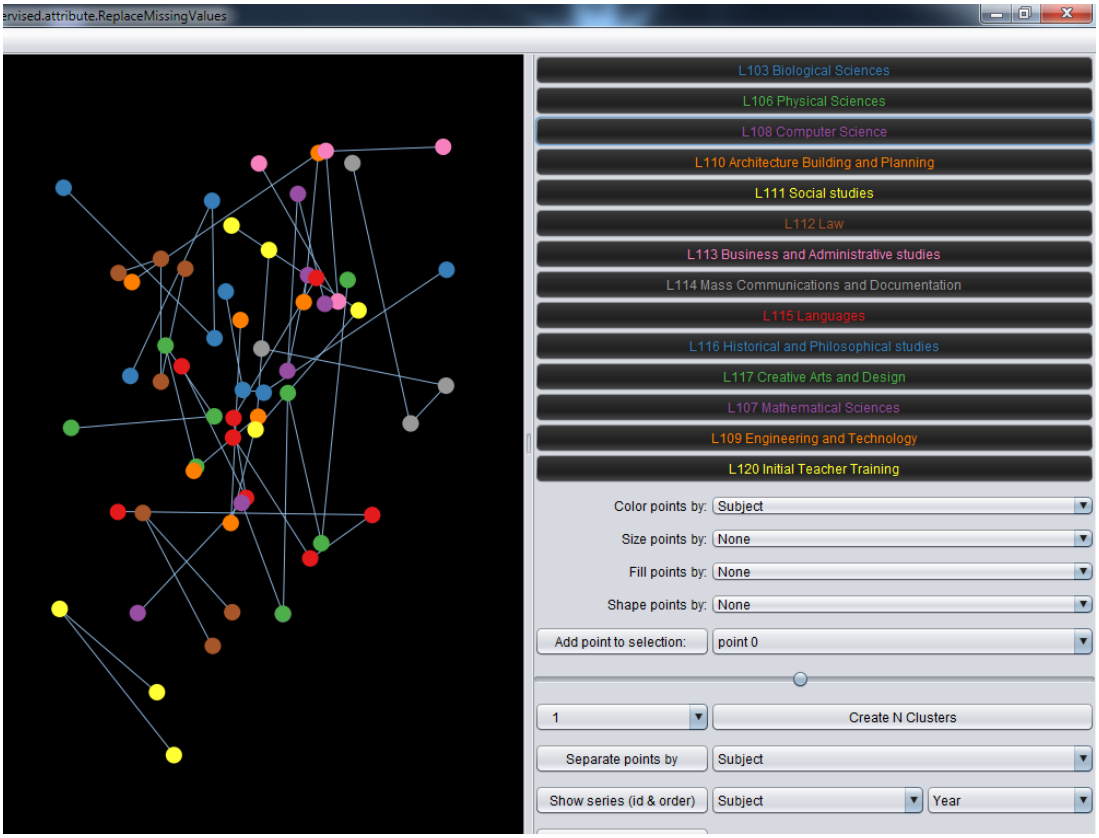


Figure 10: showing the current design of the TPP tool

8.1.1 Problems with the current user-interface

TPP in its current form has a number of problems. The way the data points are currently plotted onto the projection plane makes it difficult for users to distinguish the following items about the data:

- the year that the data point corresponds to (see note 1);
- the order of the data (see note 2) and;
- the user can only choose categorical variables rather than quantitative variables (see note 3)

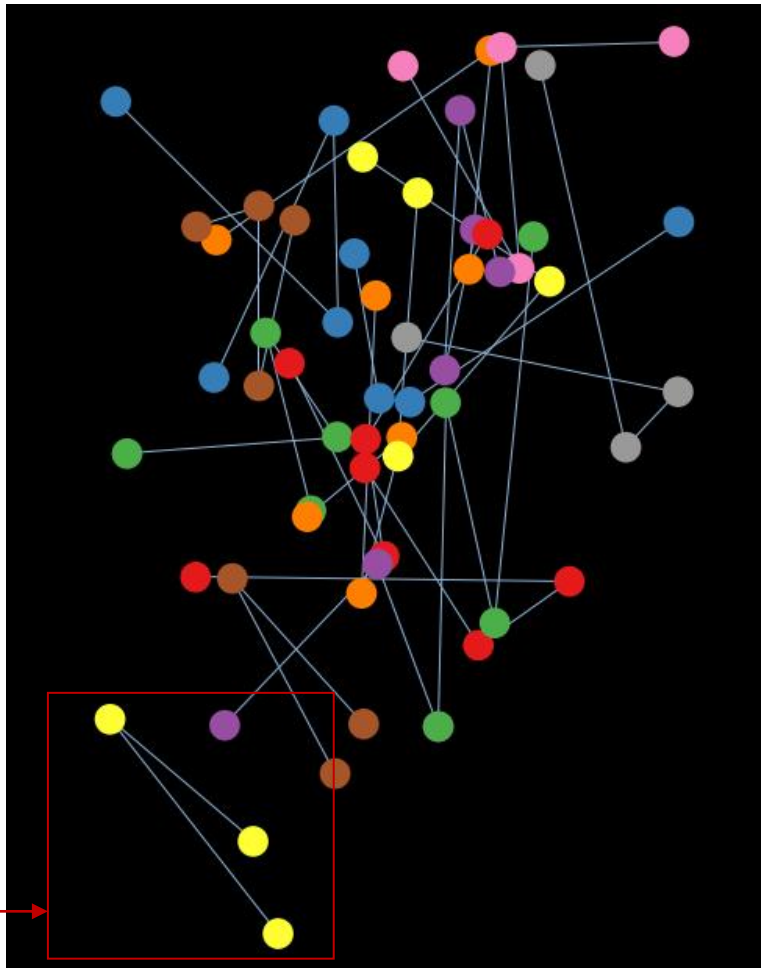
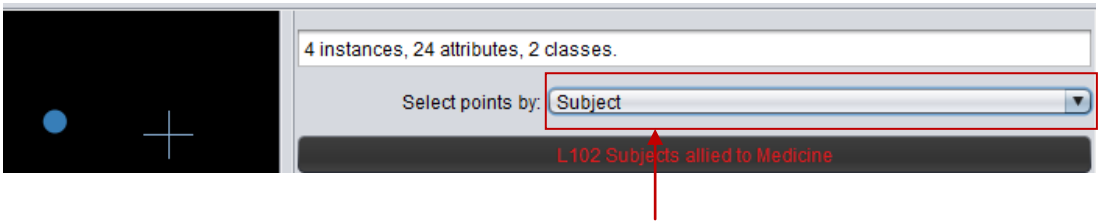


Figure 11 showing the subject identifiers plotted on the projection plane and the problems a user faces determining what year each of the points relates to

Note 1: In its current form there is no way to tell which year each of the points corresponds to. One can see that the points are related as the series feature has connected the points together with lines but currently it is unclear which year is 2008, which is 2009 and which point relates to 2010 data etc…

Note 2: Just as it is difficult to distinguish which point relates to which year for each of the subject identifiers, the same is true for the order of the data. That is to say it is difficult to determine whether or not as time passes (i.e. each years worth of data) if the NSS scores are improving or worsening.



Note 3: Currently the 'select points by' feature only allows users to select categorical variables, for example as shown above "Subject" is one of the categorical variables that can be selected. It would be useful to be able to select quantitative variables, for example select all the JACS subjects with either high or low scores for Q22

Finally, the interface is also problematic to use on screens with lower resolutions. Figure 13 below is a zoomed in view of how the interface is displayed on a screen using 1280 x 800 resolution.

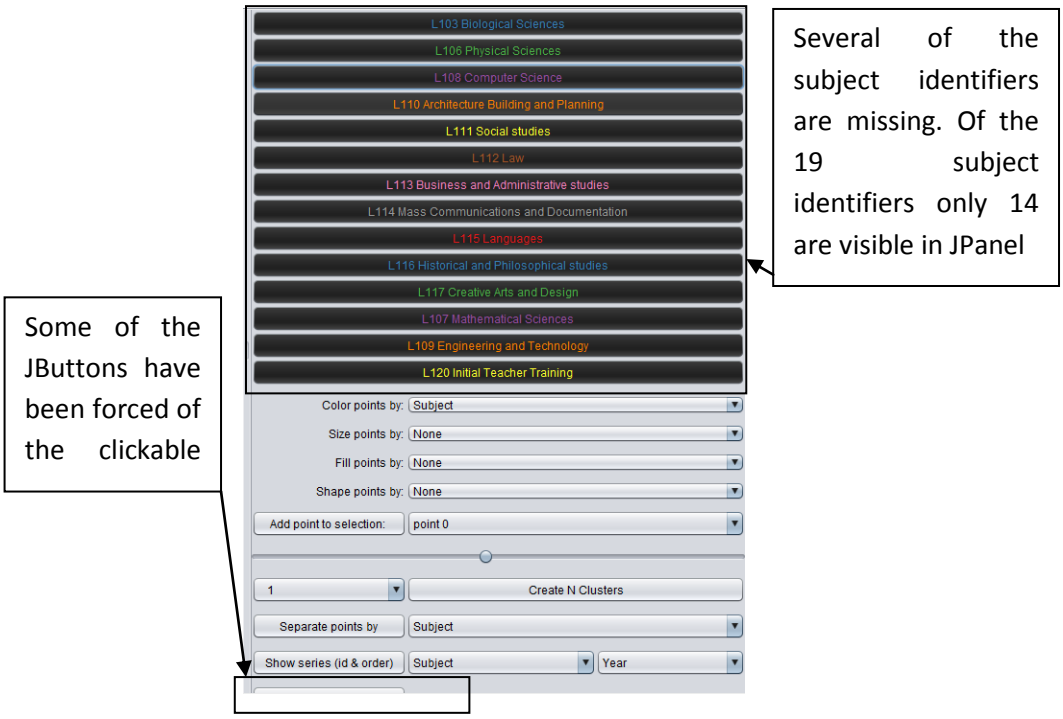


Figure 13: Showing the elements of the JPanel being forced of the screen

8.2 User requirements

Following consultation with users a set of requirements have been concluded. The user(s) require that they can:

- load data sets from a data cube
- visualise NSS scores in a graphical format
- use the resulting scatter plot (projection) to visually explore which NSS factors have the greatest effect on overall student satisfaction (i.e. Q22 of the survey) between programmes and over time

8.3 Design solutions

As discussed in section 7.1 of this chapter there are several issues that need to be addressed to make the user-interface more useable. This section provides potential solutions to solving these issues and examines each of the potential solutions and their fitness-for-purpose within the scope of this project.

8.3.1 Solving the year-order issues within TPP

There are several possible implementations that could be adopted to solving the problem of effectively visualising the year-order of the data within TPP. The following have been chosen as possibilities:

- Coloured lines
- Arrows
- Tooltips

8.3.1.1 Coloured lines solutions

Using coloured lines to show the linkage between years would be one method to help solve the year-order issue.

Using this method a key could be provided which would then used to represent a year. For example a red line could denote the link between 2007 and 2008 data points on the projection.

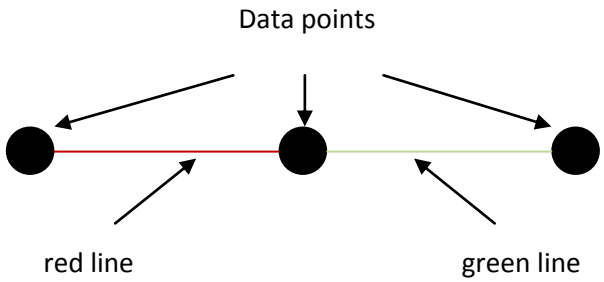


Figure 14 showing a simplified example of how a coloured line could show the links between years. A red line may represent the link between 2007 and 2008 , a green line could be used to represent the linkage between 2008 and 2009's data

Advantages of using the coloured lines solutions

Using coloured lines has some benefits:

- It can show at a glance the different links between points
- One colour can be used to represent several links across several subjects. For example red can represent the links between 2007 and 2008 data for Computer Science, Languages, Law…etc.
- Colour is a good visual stimulus and helps understanding

Disadvantages of using the coloured lines solutions

The use of coloured lines to solve the year-order problem has some problems:

- It doesn't show the order, i.e. it doesn't show at a glance which order the data is flowing. It can not show which direction leads from 2007 to 2008 and so on.
- Colours are limited and even using different shades of a particular colour could become more confusing

8.3.1.2 Arrow solution

Using lines with arrows is another method that could be used to solve the year-order problem.

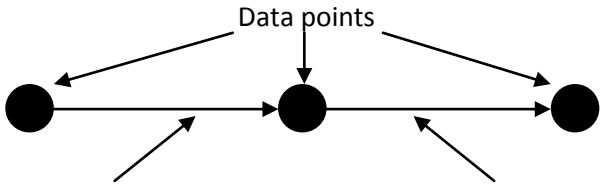


Figure 15 Arrowed lines showing the direction of flow of the data by year

Advantages of using the arrowed lines solutions

The advantage offered by using arrowed lines are as follows:

- They easily show the direction of flow between the data points on the projection plane
- Make identifying the order of the data somewhat easier

Disadvantages of using the arrowed lines solutions

The disadvantages of using arrowed lines are as follows:

- They still cannot identify the year that each of the data points correspond to
- Could clutter the projection even further and make following the flow of data more complex for the user

8.3.1.3 Tooltip solution

Tooltips are a convenient way to provide snapshot of data when the user hovers over an area. The following is an example of how tooltips could be used to help solve the problem of determining the year within TPP:

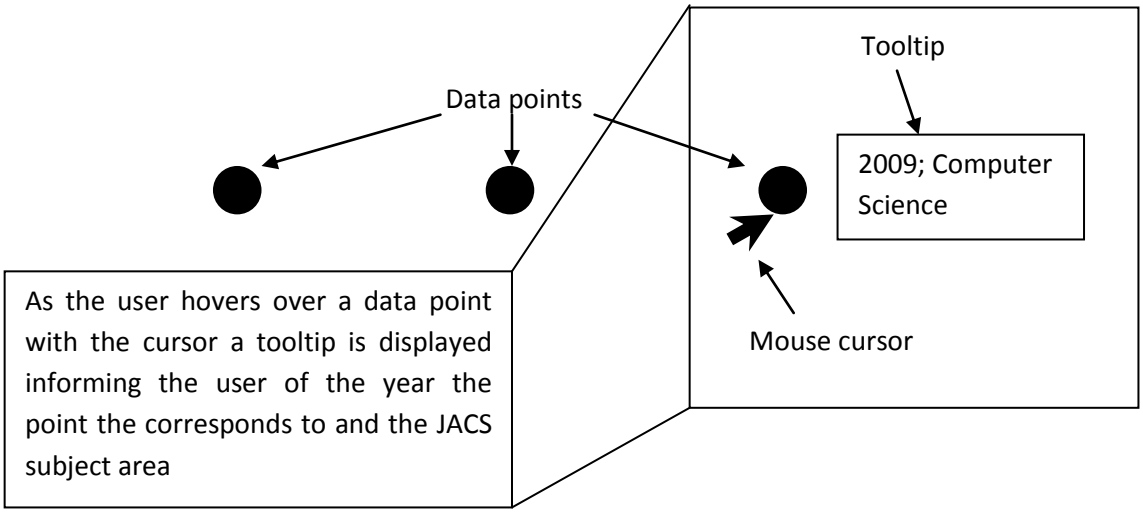


Figure 16: The proposed design showing the use of tooltips

Advantages of using tooltips

Tooltips are useful as they provide the following benefits:

- Provide a snapshot of data
- Can "label" points to allow users to quickly see what the data point represents
- Unobtrusive as they only appear when the mouse hovers over an area that is associated with a tooltip

The following table shows the criteria used to evaluate the methods available for solving the year-order issue within TPP

8.3.2 Solving the clutter issues within TPP

There are several possible implementations that could be adopted to solving the problem of clutter within TPP. The following have been chosen as possibilities:

- Tabbed JPanel
- Pop-out panels

8.3.2.1 Proposed design using tabbed panelling

The following figure (17) shows a proposed solution to solving the problem of arranging the data to ensure all the data fits on the interface regardless of the user's screen resolution.

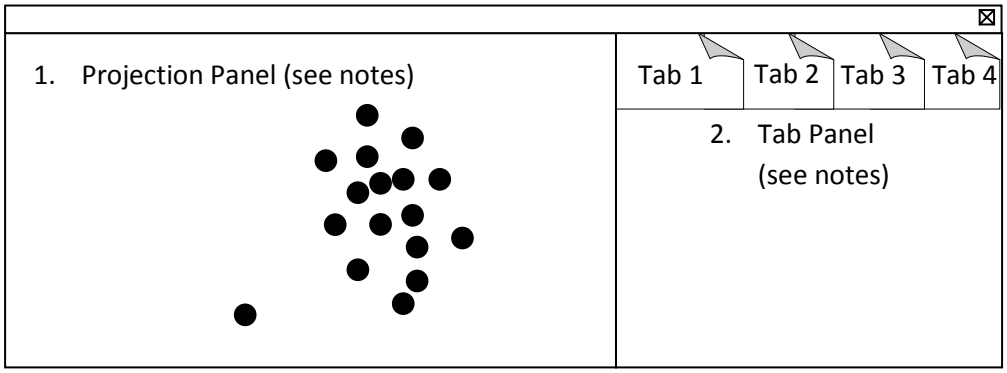
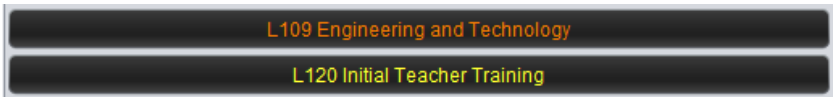


Figure 17 showing a possible solution using tabs to separate and de-clutter the JPanel on the right-hand side of the GUI

Notes:

1. Projection Panel. The projection panel is where the data is plotted.
2. Tab panel. Using the tabbed method, the current panel (a standard JPanel) would be converted to a tabbed JPanel. There would be 4 tabs, one for each of the 4 main features that are currently available on the JPanel. The panels would be split as follows:
    - A tab to deal with displaying the subject identifiers (i.e. the JACS subjects) see below for an example.



    - A second tab would handle the functions that determine the presentation of the data points. Such as the colour of the points, the size and shape etc
    - A third tab would be used to handle the series and separation functions provided by TPP.
    - A fourth tab would be used to display the projection table

Advantages of using tabbed panels

Tab panels are useful in several ways as they can be used to:

- show or hide data when necessary
- are a useful means to "de-clutter" a user-interface.
- allow for related items to be grouped together in a logical manner
- provide a natural means of separate the data
- lessen the demand placed on the user when using the user-interface

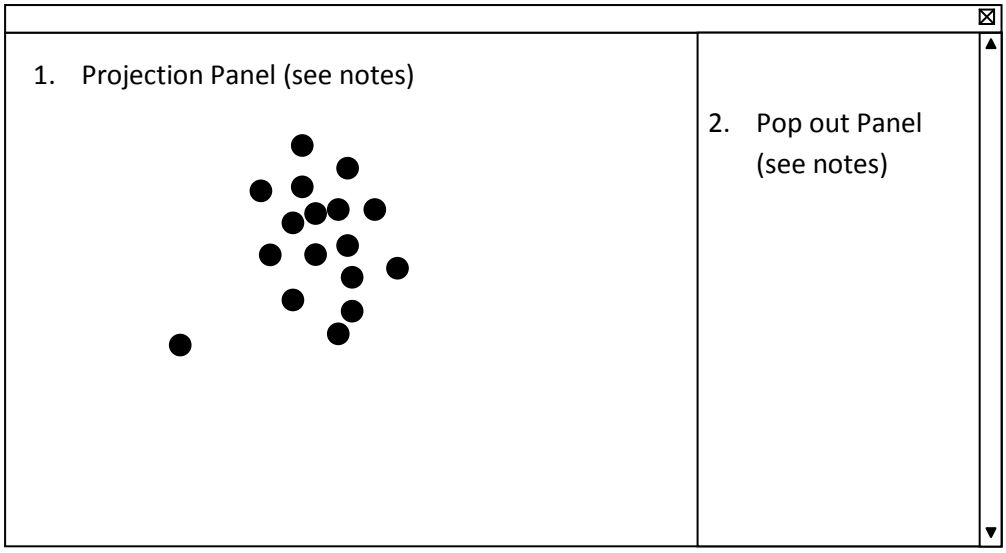
Disadvantages of using tabbed panels

Tab panels do however have some problems:

- they reach saturation point easily and add no value when there is large amounts of data
- they require much more clicking and switching by the user
- they cannot display several portions of data on the same screen and cause data to be hidden and are less transparent

8.3.2.2 Proposed design using scroll panels

The following figure shows a proposed solution to solving the problem of screen clutter using pop-out panels.



Notes:

1. Projection Panel. The projection panel is where the data is plotted.
2. Scroll panel. Using scroll panels, the user could scroll up and down to view content that was not visible on the screen. The scoll bars allow easy navigation of the panel.

Advantages of using scroll panels

Pop out panels are useful in several ways as they can be used to:

- The data remains on one panel and there's no need for the user to switch panels
- Allows unlimited amounts of data to be added to the panel
- reduces the amount of switching and clicking required by the user
- can show all the data at the same time and unlike tabs do not temporarily "hide" the data.

Disadvantages of using scroll panels

As with using tab panels, scroll panels do also have problems; however the only real issue they have is the can:

- force the user to have to scroll large amounts of data
- can encourage huge amounts of data to be squeezed into one panel, which in turn can cause confusion to the user

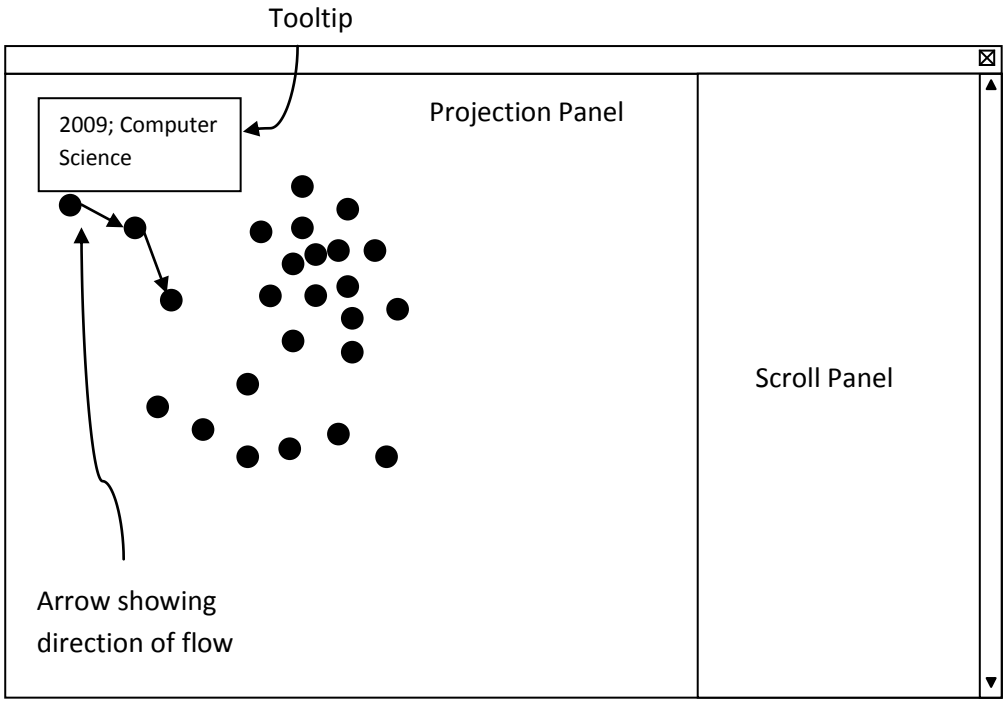
8.4 Proposed design solution

Following on from an analysis of the requirements and considering the advantages and disadvantages of each of the proposed solutions, the following design has been proposed. It combines the use of tool tips and arrowed lines along with a scroll panel.

Using a combination of tool tips and arrowed lines provides the best coverage and the solution to the problem. Furthermore using pop out panels will prove more versatile than a tabbed panel.

Additionally, the modifications will be written in Java. Java has been chosen as the language to program in as the rest of the visualisation tool has already been programmed in this language. This is a given and there has been little other consideration required as to other possible programming languages.

The proposed design is shown below.



8.6 Summary

The chapter has presented an evaluation of the current design of the visualisation tool. A new design has been proposed that employs several techniques including tooltips and an improved panel design. A list of user requirements has also been listed following on from consultation with potential users. A testing plan has also been proposed.

Chapter 9 – Hypothesis testing

This chapter outlines the results of the testing that has taken place with users and also the results of what was found from doing the actual experiment. i.e. testing the hypothesis to see if what was hypothesised was actually proven to be accurate. Firstly the hypothesis is re-stated and then the method used to test the hypothesis is explained.

9.1 Hypothesis

Targeted Projection Pursuit (TPP) can be used to visualise historical National Student Survey (NSS) data from an OLAP cube and identify which factors most influence students overall satisfaction (Q22 of the National Student Survey) between course programmes and over time.

This hypothesis was tested by carrying out the experiment as described in Chapter 5.

9.2 Method

The data used to test the hypothesis was fairly well-defined. The test set of data that was used consisted of data from the previous 4 years. The data was collated and then extracted from the data cube into an ARFF format (the format recognised by TPP).

The data was then processed in TPP and to ensure that the value of both the year and question 22 were removed from the test set (this was because we were trying to determine which of the other questions most affected the resulting score of question 22). These were our criteria for testing the rest of the data set against. Using the TPP tool, the experiment found the “best view” of the data and then produced a projection of the data.

The experiment was conducted several times with the data and then the results were recorded.

9.3 Results

The results of the experiment are shown below:

From the experiment, the following questions were found to have the most significant affect on the result for question 22 of the National Student Survey.

Question Number	Question
10	I have received sufficient advice and support with my studies.
11	I have been able to contact staff when I needed to.
6	Assessment arrangements and marking have been fair.
17	I have been able to access general IT resources when I needed to.
16	The library resources and services are good enough for my needs.
20	My communication skills have improved.
12	Good advice was available when I needed to make study choices.
14	Any changes in the course or teaching have been communicated effectively.
18	I have been able to access specialised equipment, facilities or room when I needed to.
1	Staff are good at explaining things.
3	Staff are enthusiastic about what they are teaching.
2	Staff have made the subject interesting.
7	Feedback on my work has been prompt.
5	The criteria used in marking have been clear in advance.
4	The course is intellectually stimulating.
9	Feedback on my work has helped me clarify things I did not understand.
21	As a result of the course, I feel confident in tackling unfamiliar problems.
19	The course has helped me present myself with confidence.
13	The timetable works efficiently as far as my activities are concerned.
8	I have received detailed comments on my work.
15	The course is well organised and is running smoothly.

Most influential

Least influential

From the results of our testing it has been found that the factors which most affect the result achieved for question 22 from the National Student Survey. In the testing, question 10 was found to have the most significance, in other words receiving sufficient advice and support with their studies is what has the most significance for achieving overall satisfaction (i.e. achieving good results for Q22).

What is interesting to discover from the results is that the second and third most important factors (question 11 and question 6) are both questions which aim to gauge opinions about similar areas of the students experience; question 11- "I have been able to contact staff when I needed to" and question 6 – "Assessment arrangements and marking have been fair" – both straddle the categories of assessment and feedback and academic support. This suggests that there is a trend in student views. In other words the most influential factors affecting overall student satisfaction are with regard to support and feedback.

These results seems to be slightly differing to what Langan et al (2010) claim in their research – *Extracting useful information from the UK's National Student (Satisfaction) Survey* (2010) – who claimed that actually enhancing the feedback and the process of providing feedback does not directly enhance the satisfaction achieved for question 22. They also claimed that for the year 2009 the five most influential questions in order of influence are: question 15, 4, 1, 21, and 10 (Langan et al, 2010). Only one of their questions – question 10 – was included in both their results set and ours.

This is interesting and would certainly warrant further investigation to find out these differences in results and to discover the differences in their research experiment to try and discover possible reasons for these disagreements in the results set.

It should however be noted that the results that Langan et al (2010) came up with based on data from Manchester Metropolitan University only. Perhaps and experiment that used their data would produce more similar results. It would also be interesting to see whether what they are saying is the most influential

results at their institution are the most influential factors across the country as a whole. Of course the same limitations were to be applied to our own experiment which simply looks only at Northumbria University's results. Perhaps a larger-scale experiment may provide more substantial results.

Finally, comparing results with other researchers work may have been able to provide a more accurate picture of the results. Comparing it to Langan et al's (2010) work only perhaps somewhat limits the scope of this. However at the time of carrying out the research their work was the only comparable results available and given the small data set size the results can only provide a small snapshot of the general trends for the previous 4 years. A larger, more comprehensive data set if it had been available would perhaps have provided us with more substantial results. However this was not possible given the short time the NSS survey has been in operation.

9.4 User evaluation

After the tool was modified and the experimental research had been carried out there was a series of evaluations carried out with domain experts from the University, these were users who conduct analysis on the National Student Survey results for the University.

The following conclusions and information was gathered from the users and the users commented that:

- The tool is useful for providing a global picture of the Institution as a whole. That is to say that the tool can provide a overview of each of the Schools with in the University in a quick and efficient manner and is an improvement on the current process
- It has better graphical value than the bar graphs currently used
- It would be a useful tool for senior management within the University
- It would be a useful tool for comparing results at a School level; by this they mean individual schools analysing their own results over time to understand the progress their individual school is making rather than at an institutional level

- The time element (i.e. data for each of the years of the survey) is very informative and provides a simpler and more intuitive means of comparison

The following suggestions were made to improve the tool:

- Provide a user guide for the tool
- Compare data against other institutions at both subject and institutional level; that is to say allow users to compare subjects e.g. Mathematics at Northumbria University with the results for Mathematics at Newcastle University.
- Combine the results of the University's internal survey which asks students to answer the same questions as the National Student Survey. Then by combing the results of both surveys see whether or not there is any difference in the results
- Compare the results at a Programme level
- Rather than focussing on just question 22 of the survey, perhaps selecting some of the other questions highlighted in the research by Langan et al (2010) and seeing if this affected the result in any way.

9.5 Design fitness for purpose

In the previous chapter, a proposed improved design was put forward that included the use of tool tips and improved panel design, along with the use of arrows to provide improved understanding.

Most of these requirements have been met; the panel design has been improved and arrows were implemented. However the use of tool tips has not been met. This is due to two reasons. Tool tips were more difficult to implement than first thought. They proved difficult to implement using TPP's current program structure and time for the research was limited. Given this lack of time it was decided not to implement tool tips at this stage, although it is still seen as a useful addition to the TPP user interface.

9.6 Summary

In this chapter there has been an explanation of the hypothesis testing including the method, the results and some feedback from the user evaluations. There is some analysis of the results gained from the experiment and considers their significance and reliability and compares the results to a similar experiment carried out by Langan et al (2010) and considers whether or not the design criteria have been met or not.

Chapter 10 – Evaluation

10. Introduction

In this chapter there is an evaluation of the project as a whole. It evaluates the research process (the literature review) and considers the possibility that there may have been other techniques or methods available which were overlooked. It also considers whether or not other hypotheses could have been tested, as well as other methods that could have been used to evaluate the visualisation tool, the results and their reliability and finally considers whether or not the users' requirements have been met.

10.1 Evaluation of the Research Process

Given the scope of the project the algorithms, tools and techniques that could have been researched were limited. OLAP visualisation is a relatively unexplored field so the number of research resources available was quite limited. However, storing the data in OLAP cubes was the most appropriate method for storing the data. Of course there are several other methods that could have been researched or implemented. The data could simply have been stored in a relational database such as Microsoft Access and then convert the data into a readable format that the visualisation tool could read, for example a the database could have been saved in a csv format (Microsoft, 2010).

Furthermore, with regards to the literature survey, there may have been other techniques that could have been researched rather than focussing only on linear and non-linear methods. However, preliminary research suggested using either one of these two methods would be the best route to follow and that these are the most popular. It was also apparent from an early stage that finding a tool that was already pre-built would be more useful than building a tool from scratch. The identification of the TPP algorithm and tool was made at an early stage appears to have been a good choice. It was also useful because it was compatible with the other tools (WEKA and Mondrian) which made the conducting the research experiment easier.

Additionally the choice of TPP made sense as it was proven to be the most effective method for visualising data in this way. Perhaps it would have been useful to have tried several methods to see what the differences in results, if there were any. Undoubtedly there will be other techniques available that could have been tested and reviewed. However on reflection, the best spread of techniques were reviewed in the literature although it would have been useful to have researched further some other techniques if time had permitted.

10.2 Alternative hypotheses

The scope of the hypothesis was limited somewhat by the client's requirements. They were most interested in overall satisfaction and therefore the hypothesis had to have reference somewhere in it to question 22 of the survey.

There was however the question of Northumbria University's own internal survey which aims to mirror the National Student Survey as closely as possible. There may have been scope to test a hypothesis that looked at whether or not the same factors influenced overall satisfaction in the internal survey as the ones that influenced satisfaction in the National Student Survey. This would be an interesting hypothesis to test as any differences in results would make for an interesting investigation in to why the results were different. Finding out how what people were saying in one was different from what they were saying in the other survey.

Overall though it is thought the most suitable hypothesis was chosen both for research value to the client and in terms of investigative purposes as a piece of research work.

10.3 Experimental results and conclusions

It was interesting to discover that the results in this experiment were different from the results found from the research that Langan et al (2010) carried out. As has been discussed in the previous chapter, there are clear reasons why this may have been apparent. Both our experiment and theirs was based on results for a single-institution. It has to be noted that no institution will be the same as another one and therefore the results cannot be applied in general terms.

Additionally, the amount of survey data available is relatively small. Four years worth of data is a relatively small sample of data and an anomaly in one years results could potentially cause some issues in accurately predicting the true factors that affect the results. There may be results for a particular year that for whatever result are abnormal. Problems that normally wouldn't happen that for some reason did happen in a particular year in the data set could artificially affect the results in such a small data set. With this in mind it would be most useful to collate more data when it becomes available to smooth out and remove any particular "noise" or abnormal results and then a more accurate picture may be gained.

10.4 User requirements evaluation

The user requirements were evaluated and checked to see whether they met the clients' requirements. The requirements of the user were to be able to:

- load data sets from a data cube. This has been met. The user is able to load an ARFF file produced by Mondrian and using the TPP tool can load the resulting data set into TPP
- visualise NSS scores in a graphical format. This requirement has been met and the user is able to use TPP to provide them with a graphical visualisation of the data.
- use the resulting scatter plot (projection) to visually explore which NSS factors have the greatest effect on overall student satisfaction (i.e. Q22 of the survey) between programmes and over time. This requirement has been satisfied and the user is able to use TPP to explore and discover which of the other 21 questions has the greatest influence on overall student satisfaction (question 22).

10.5 Evaluation criteria

The criteria used to evaluate the experiment were relatively limited. The best way to evaluate the experiment was to ask the client to provide feedback. Asking the client whether or not what had been done was what the wanted and

asking for their feedback was the best course of action. They were the client so asking them to provide feedback was the only sensible means of testing what we had done. Perhaps some more user evaluation sessions would have been useful and consultation with more users would have been useful, but generally, the domain experts who evaluated the work and the criteria used to evaluate were the best means for testing our work.

10.6 Summary

In this chapter we have evaluated aspects of the project including the effectiveness of the literature review and whether there were any other techniques that could have been used instead. It also presented any alternative hypotheses and justifies why the particular hypothesis was chosen. Furthermore, the user requirements have been evaluated along with the criteria used to evaluate the tool and draws conclusions from the results and their reliability and effectiveness.

Chapter 11 – Conclusions and Recommendations

11. Introduction

This chapter outlines the conclusions that have been drawn from this research project. Furthermore, there are several recommendations for further work in the future.

11.1 Conclusions

During the project, the aim was to produce a method for visualising National Student Survey data that would help aid users in their understanding of the data. Achieving this was of great importance as it has potential to save time and effort in the future for users.

Throughout the development process, it has been concluded that choosing Targeted Projection Pursuit (TPP) was an extremely convenient, efficient, accurate and useful tool for visualising multidimensional data. This is due to the architecture on which Targeted Projection Pursuit has been developed. The fact that TPP is built upon the WEKA toolkit proved extremely useful as it made the transfer of data from the data cube into TPP was seamless due to the OLAP cube builder also being built upon the same WEKA platform.

The tool, when the data was loaded into it was able to provide accurate identification of which factors from the National Student Survey most affected the score for overall satisfaction in the survey, therefore answering our original research question and hypothesis.

11.2 Recommendations for future work

There are several recommendations that could be suggested to improve the software and overall process of visualising the data.

One such recommendation for further enhancement in the future would be developing a way to connect the Targeted Projection Pursuit Tool and Mondrian in a more direct way. Currently the OLAP cube builder (Mondrian) produces an ARFF file which the users then have to load from Targeted Projection Pursuit. It would be desirable to have way to allow a user in one-click to load a data cube

from Mondrian and then have TPP convert it to an ARFF format seamlessly. This could be achieved by creating a new Java class in TPP which would exploit the underlying architecture of Mondrian which could then trigger the function in Mondrian that converts a cube into ARFF format. After this, then the load function could be actioned in TPP and would save both time and effort.

In terms of research work into the actual field of visualisation of such data, then there is huge scope for more work. Extraction of data from an OLAP cube and then visualising it further is very limited as was pointed out in the literature review chapter. Researching and exploiting this area would be potentially very worthwhile.

Furthermore, following on from discussions with the client, a number of possible avenues for future work arose. It would be useful to combine the results set from the National Student Survey with the results of Northumbria University's own internal survey. Combining the results and/or analysing the results side-by-side would make for an interesting piece of research work. Finding out whether what people answered in one of the survey's reflected what they were saying in another. Using the results form the internal survey would also provide a means for analysing the data at a programme level rather than at a school level, something which our users seem keen to know more about, so taking the internal survey results and visualising them would potentially be a worthwhile piece of research to conduct.

Additionally, user feedback suggests that perhaps comparing the NSS results against other institutions from across the UK would be useful at gaining an overall picture of how the university stands amongst it peers and could provide some potential for providing the Institution with ideas that they could use to boost overall student satisfaction for question 22 of the NSS survey.

Finally, further improvements to the user-interface of the TPP tool would be useful. Extending the features that were not achieved in this project, such as the tool tips would make the tool particularly useful. Implementing the tool tip

function that was unable to be implemented due to time constraints on the project would be one of these improvements that could be made in the future.

11.3 Summary

This chapter has provided a list of conclusions that have been reached throughout the project and has offered several suggestions for possible future work that could be conducted and this concludes the report.

REFERENCES

Ammoura, A., Zaiane, O.R., & Ji, Y. (2001). Towards a Novel OLAP Interface to Distributed Data Warehouses. Proceedings of the 3rd International Conference on Data Warehousing and Knowledge Discovery, LNCS Vol. 2114, 174-185

BI-LITE (no date*) BI-Lite - FREE Microsoft SQL Server OLAP Cube Building Tool*. [Online]. Available at: http://www.bi-lite.co.uk/product/CUBEITZEROEnquiry.aspx (Accessed: 15 November 2010).

Borg, I. and Groenen, P (2005) *Modern Multidimensional Scaling: theory and applications*. 2nd edn. New York: Springer-Verlag.

Borgatti, S. P. (1997) *Multidimensional Scaling*. [Online]. Available at: http://www.analytictech.com/borgatti/mds.htm (Accessed: 23 October 2010).

Cheng, L. T., Kerr, B. (no date) *Project: Bloom.* Cambridge, MA: IBM Watson Research Center [Online]. Available at: http://domino.watson.ibm.com/cambridge/research.nsf/2422527657l0c19485256979004d289c/864fa36402234708852570f90079a47a?OpenDocument (Accessed: 22 March 2010).

Codd, E.F., Codd, S. B. and Salley, C.T. (1992). *Providing OLAP (on-line analytic processing) to user-analysts: an IT mandate*. [Online] Available at: http://www.minet.uni-jena.de/dbis/lehre/ss2005/sem_dwh/lit/Cod93.pdf

Cox, T. F. and Cox, M. A. A. (2001) *Multidimensional Scaling*. 2nd edition. Chapman and Hall.

Cuzzocrea, A. & Mannsaman, S. (2009) 'Models, Issues, and Techniques in OLAP Visualization', OLAP Visualisation, pp. 1439 - 1446 [Online]. Available at: http://www.inf.uni-konstanz.de/dbis/publications/download/CM-EDWM08.pdf (Accessed: 10 December 2010).

Dubler, C. and Wilcox, C. (2002) *Just What Are Cubes Anyway? (A Painless Introduction to OLAP Technology).* Denver, CO: Microsoft Corporation [Online]. Available at: http://msdn.microsoft.com/en-us/library/aa140038(office.10).aspx (Accessed: 17 March 2010).
Enshaie, A. and Faith, J. (2010) 'Feature Selection with Targeted Projection Pursuit'. (Unpublished)

Faith, J. (2007)' Targeted Projection Pursuit for Interactive Exploration of High-Dimensional Data Sets' *11th International Conference Information Visualization* Zürich, Switzerland 2 - 6 July 2007. IEEE Computer Society

REFERENCES

Freidman, V. (2008) *Data Visualization and Infographics*. [Online]. Available at: http://www.smashingmagazine.com/2008/01/14/monday-inspiration-data-visualization-and-infographics/ (Accessed: 20 October 2010).

Friedman, J. H. and Tukey, J. W. (1973) 'A Projection Pursuit Algorithm for Exploratory Data Analysis', *IEEE Transactions on Computer*, 23 (9), pp. 881-890 [Online]. Available at: http://portal.acm.org/citation.cfm?id=1311448 (Accessed: 11 October 2010).

Friendly, M. and Denis, D. J. (2001) *Milestones in the history of thematic cartography, statistical graphics, and data visualization*. [Online]. Available at: http://www.math.yorku.ca/SCS/Gallery/milestone/ (Accessed: 19 December 2010).

Gebhardt, M., Jarke, M., & Jacobs, S. (1997) ACM International Conference on Management of Data. Tucson, Arizona, USA May 11 - 15. ACM.

Hanrahan, P., Stolte, C., & Mackinlay, J. (2007). *Visual Analysis for Everyone: Understanding Data Exploration and Visualization*. (Accessed: 27 October 2010).

HEIDI (2010) *heidi*. [Online]. Available at: https://heidi.hesa.ac.uk/ (Accessed: 19 September 2010).

Huber, P. J. (1985) 'Projection Pursuit', *The Annuals of Statistics*, 13 (2), pp. 435–475 [Online]. Available at: http://www.stat.rutgers.edu/~rebecka/Stat687/huber.pdf (Accessed: 27 October 2010).

Iles, S. (2010) Email to Joseph Hogg, Various dates in 2010.

Iles, S. (2011) Conversation with Joseph Hogg, 7 January 2011 .

Ipsos MORI. (2010) *The National Student Survey.* [Online]. Available at: http://www.thestudentsurvey.com/faqs.asp (Accessed: 22 March 2010).

Jolliffe, I. T. (2002) Principal Component Analysis. 2nd edn. New York: Springer.

Lang, S. (1987) Linear algebra. Berlin: Springer-Verlag.

REFERENCES

Langan, M., Fielding, A. and Dunleavy, P. (2010) *Extracting useful information from the UK's National Student (Satisfaction) Survey*. Manchester Metropolitan University (22-23 June).

Likert, R. (1932) 'A technique for the measurement of attitudes', Archives of Psychology, 22 (140), 1932, pp. 1 - 55 [Online]. Available at: http://www.citeulike.org/user/robertlischke/article/2731047 (Accessed: 22 September 2010).

Mailvaganam, H. (2007) *Introduction to OLAP.* [Online]. Available at: http://www.dwreview.com/OLAP/Introduction_OLAP.html (Accessed: 16 March 2010).

Maniatis, A. S., Vassiliadis , P., Skiadopoulos, S. and Vassiliou, Y. (2003) ' Advanced Visualization for OLAP' *6th ACM international workshop on Data warehousing and OLAP* New Orleans, LA, USA 7 November 2003. New York, NY, USA: ACM.

Njoroge, J. (2010) Conversation with Joseph Hogg, Various dates.

Njoroge, J. (2010) Email to Joseph Hogg, Various dates in 2010.
Pearson, K. (1901) 'On lines and planes of closest fit to systems of points in space', *Philosophical Magazine*, 2 (6), pp. 559 - 572 [Online]. Available at: http://stat.smmu.edu.cn/history/pearson1901.pdf (Accessed: 17 October 2010).

Pentaho Corp. *Open source analysis OLAP server written in Java. Enabling interactive analysis of very large datasets stored in SQL databases without writing SQL. | Mondrian: Pentaho Analysis*. [Online]. Available at: http://mondrian.pentaho.com/ (Accessed: 13 October 2010).
Phelps, N. (2010) Email to Joseph Hogg, 24 March 2010.
Post, H. F., Nielson, G. M. & Bonneau, G. P. (2003*) Data Visualization: The State of the Art*. Dordrecht: Kluewer.

Ramsden, P. (1991) 'A Performance Indicator of Teaching Quality in Higher Education: The Course Experience Questionnaire.', Studies in Higher Education, 16 (2), 1991, pp. 129 - 150 [Online]. Available at: http://eric.ed.gov (Accessed: 14 September, 2010).

Sifer, M. (2006) 'User Interfaces for the Exploration of Hierarchical Multi-dimensional Data', Baltimore. MD, USA 31 October - 2 November. IEEE, pp. 175 - 182. 19 October 2010).

REFERENCES

Smith, A. (no date) *Developments in Data Visualisation*. [Online]. Available at: www.oecd.org/dataoecd/5/40/38188055.pdf (Accessed: 3 January 2011).

Strachan, R. (2011) Conversation with Joseph Hogg, Various dates in 2010.

Techapichetvanich, K., & Datta, A. (2005). Interactive Visualization for OLAP. Proceedings of the International Conference on Computational Science and its Applications (Part III), 206-214

Tegarden, D.P. (1999). Business Information Visualization. Communications of the AIS, 1(1), Article 4

The Good University Guide. *University League Table Methodology*. [Online]. Available at: http://www.thecompleteuniversityguide.co.uk/single.htm?ipg=8728 (Accessed: 23 September 2010).

The Guardian. *University guide 2011: University league table | Education | guardian.co.uk*. [Online]. Available at: http://www.guardian.co.uk/education/table/2010/jun/04/university-league-table (Accessed: 23 September 2010).

The University of Alberta. (1999) *The University of Alberta* [Online]. Available at: http://www.ualberta.ca/ (Accessed: 17 December 2010).

Times Online. (2008) *University Rankings League Table | The Sunday Times University Guide 2010 - Times Online*. [Online]. Available at: http://extras.timesonline.co.uk/stug/universityguide.php (Accessed: 23 September 210).

University Alliance. (2010) *University Alliance*. [Online]. Available at: http://www.university-alliance.ac.uk/ (Accessed: 22 December 2010).

University of Waikato. (no date) *Weka 3 - Data Mining with Open Source Machine Learning Software in Java*. [Online]. Available at: http://www.cs.waikato.ac.nz/ml/weka/ (Accessed: 19 October 2010).
VisuMap (no date) *What is PCA?* [Online]. Available at: http://www.visumap.net/Resources/Demo/WhatIsPca.htm (Accessed: 23 November 2010).

Weskamp, M. (2004) *Newsmap.* [Online]. Available at: http://marumushi.com/projects/newsmap (Accessed: 23 March 2010).

REFERENCES

Wilkinson, L. (2005). *The Grammar of Graphics*. 2nd edn. Chicago: Springer.

Yang, K., Li, Y., Luo, Q., Sander, P. V. ands Shi, J. (2009) 'I3DC: Interactive Three-Dimensional Cube', *25th IEEE International Conference on Data Engineering, ICDE 2009*. Shanghai, China 29 March - 2 April 2009. Inst. of Elec. and Elec. Eng. Computer Society, pp. 1475-1478. [Online]. Available at: http://ieeexplore.ieee.org.wf2dnvr2.webfeat.org/stamp/stamp.jsp?tp=&arnumber=4812551 (Accessed: 17 March 2010).

Zaïane, O. R. (1999) *Department of Computing Science: CMPUT 690: Principales KDD (Glossary).* [Online]. Available at: http://webdocs.cs.ualberta.ca/~zaiane/courses/cmput690/glossary.html (Accessed: 23 March 2010).

Dubler, C. and Wilcox, C. (2002) *Just What Are Cubes Anyway? (A Painless Introduction to OLAP Technology).* Denver, CO: Microsoft Corporation [Online]. Available at: http://msdn.microsoft.com/en-us/library/aa140038(office.10).aspx (Accessed: 17 March 2010).

Faith, J. (2007)' Targeted Projection Pursuit for Interactive Exploration of High-Dimensional Data Sets' *11th International Conference Information Visualization* Zürich, Switzerland 2 - 6 July 2007. IEEE Computer Society

Friendly (2008) *Milestones in the history of thematic cartography, statistical graphics, and data visualization.* York, Ontario: York University [Online]. Available at: http://www.math.yorku.ca/SCS/Gallery/milestone/milestone.pdf (Accessed: 22 March 2010).

Ipsos MORI. (2010) *The National Student Survey.* [Online]. Available at: http://www.thestudentsurvey.com/faqs.asp (Accessed: 22 March 2010).

Mailvaganam, H. (2007) *Introduction to OLAP.* [Online]. Available at: http://www.dwreview.com/OLAP/Introduction_OLAP.html (Accessed: 16 March 2010).

Maniatis, A. S., Vassiliadis , P., Skiadopoulos, S. and Vassiliou, Y. (2003) ' Advanced Visualization for OLAP' *6th ACM international workshop on Data warehousing and OLAP* New Orleans, LA, USA 7 November 2003. New York, NY, USA: ACM.

Techapichetvanich, K. and Datta, A. (2005) 'Computational Science and Its Applications – ICCSA 2005', *Lecture Notes in Computer Science,* 3482/2005, 2 May 2005, pp. 206 - 214 *Computer Graphics and Rendering Workshop* [Online]. Available at: http://portal.acm.org/citation.cfm?id=956063 (Accessed: 20 March 2010).

Zaïane, O. R. (1999) *Department of Computing Science: CMPUT 690: Principales KDD (Glossary).* [Online]. Available at:

REFERENCES

http://webdocs.cs.ualberta.ca/~zaiane/courses/cmput690/glossary.html (Accessed: 23 March 2010).

Weskamp, M. (2004) *Newsmap.* [Online]. Available at: http://marumushi.com/projects/newsmap (Accessed: 23 March 2010).

Yang, K., Li, Y., Luo, Q., Sander, P. V. ands Shi, J. (2009) 'I3DC: Interactive Three-Dimensional Cube', *25th IEEE International Conference on Data Engineering, ICDE 2009*. Shanghai, China 29 March - 2 April 2009. Inst. of Elec. and Elec. Eng. Computer Society, pp. 1475-1478. [Online]. Available at: http://ieeexplore.ieee.org.wf2dnvr2.webfeat.org/stamp/stamp.jsp?tp=&arnumber=4812551 (Accessed: 17 March 2010).

APPENDIX A

TERMS OF REFERENCE FOR THE DISCIPLINE OF MSC COMPUTER SCIENCE AT THE UNIVERSITY OF NORTHUMBRIA AT NEWCASTLE

Joseph Hogg 05006060
MSc Computer Science

Project Supervisor: Joe Faith
TOR Reviewer: Ian Bradley
Second Marker: Martin Wonders

APPENDIX A

Applying Data Visualisation Techniques to Data Sets Derived from OLAP Cubes Used to Store Data from the National Student Survey

Background to the Project

Visualising high-dimensional data sets, by their very nature, can be difficult to represent particularly when there are many dimensions to the data. There comes yet another issue when one tries to take data from an Online Analytical Processing (OLAP) cube (sometimes referred to as a data cube or a hypercube) and then wishes to visualise the data in a more meaningful and clearer format. Finding ways to effectively use, extract and display the data is a major problem with much work being undertaken to try and improve and enhance the current techniques. It is thought that by using visualisation techniques goes some way to improving and solving this issue.

Data Visualisation

Data visualisation is, "the science of visual representation of data" (Friendly, 2009). Thus, data visualisation is when data has been extracted and represented in some schematic form and displayed by representing all of a data's attributes and variables. The University of Alberta (1999) defines data visualisation as the "visual interpretation of complex relationships in multidimensional data"

The Bloom Diagram is an example of a data visualisation tool and allows users to visualise the contributions of particular individuals to open source projects (Cheng & Kerr; IBM, no date). Another notable example of data visualisation in action is *Newsmap*, which sizes news stories based on their popularity and clusters similar stories to make patterns more recognisable (Weskamp, 2004).

The main purpose of data visualisation is to communicate information in a clearer and much more effective way by using graphical representations and as such is a form of visual modelling.

APPENDIX A

Figure 1: The Bloom Diagram

OLAP Cubes

An OLAP cube is in essence, a de-normalised database. Their multi-dimensional structure - regularly containing more than 20 dimensions - allows for fast analysis of data and for them to be searched quickly effectively and efficiently. However, they have a fundamental flaw, they are difficult to use to represent data in a manner that is useful or worthwhile a essentially they are simply a storage facility. Furthermore, they present problems by prohibiting users from easily drilling through the cube to effectively mine the data they want (Kobielus, 2008).

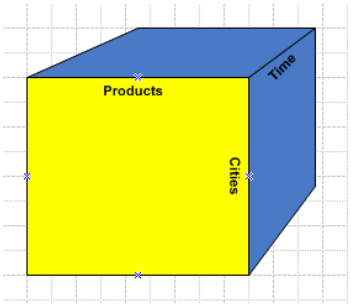


Figure 2: A simple representation of an OLAP Cube with 3 Dimensions

(Roeder, 2008)

OLAP cubes, according to Mailvaganam (2007), give users the advantage of being able to "slice and dice at will" and allow users to compare data that

ordinarily would normally not be related to one another or would otherwise be incompatible with one another. Typical operations that can be performed on an OLAP cube in Drill-down, Drill-through, Roll-up, Slice, Dice and Pivot (University of Alberta, 1999).

National Student Survey

The National Student Survey, according to its organisers is a national imitative which,

> "*...asks final year undergraduates and students in their final year of a course leading to undergraduate credits or qualifications to provide feedback on their courses in a nationally recognised format*"
>
> (National Student Survey, 2010)

The survey consists of 22 questions that cover 7 areas, of which the results of the questions are made available to a student's institution and student's union. The answers, which are anonymous, are then used by an institution, "to facilitate best practice and to enhance the student learning experience," (National Student Survey, 2010).

Some universities may also require additional questions be asked of their students, but all students are required to answer the basic 22 questions set out by the National Student Survey. Northumbria University is one of these institutions and requires its students to answer several more questions.

Current situation

Currently the "client", Northumbria University - through their Corporate Planning Department – use OLAP cubes to store data taken from the National Student Survey (NSS) and a system known as "HEIDI," then use the output from the OLAP cube to make predictions about what targets they should be meeting both in the long and short-term as well as allowing them to reflect on where they have come from and if improvements are being made as expected in a variety of areas.

APPENDIX A

The data they get from the OLAP cubes results in a Taylor Square being produced and data from it is then used to produce graphs, charts and other visual aids that attempt to represent the data more effectively.

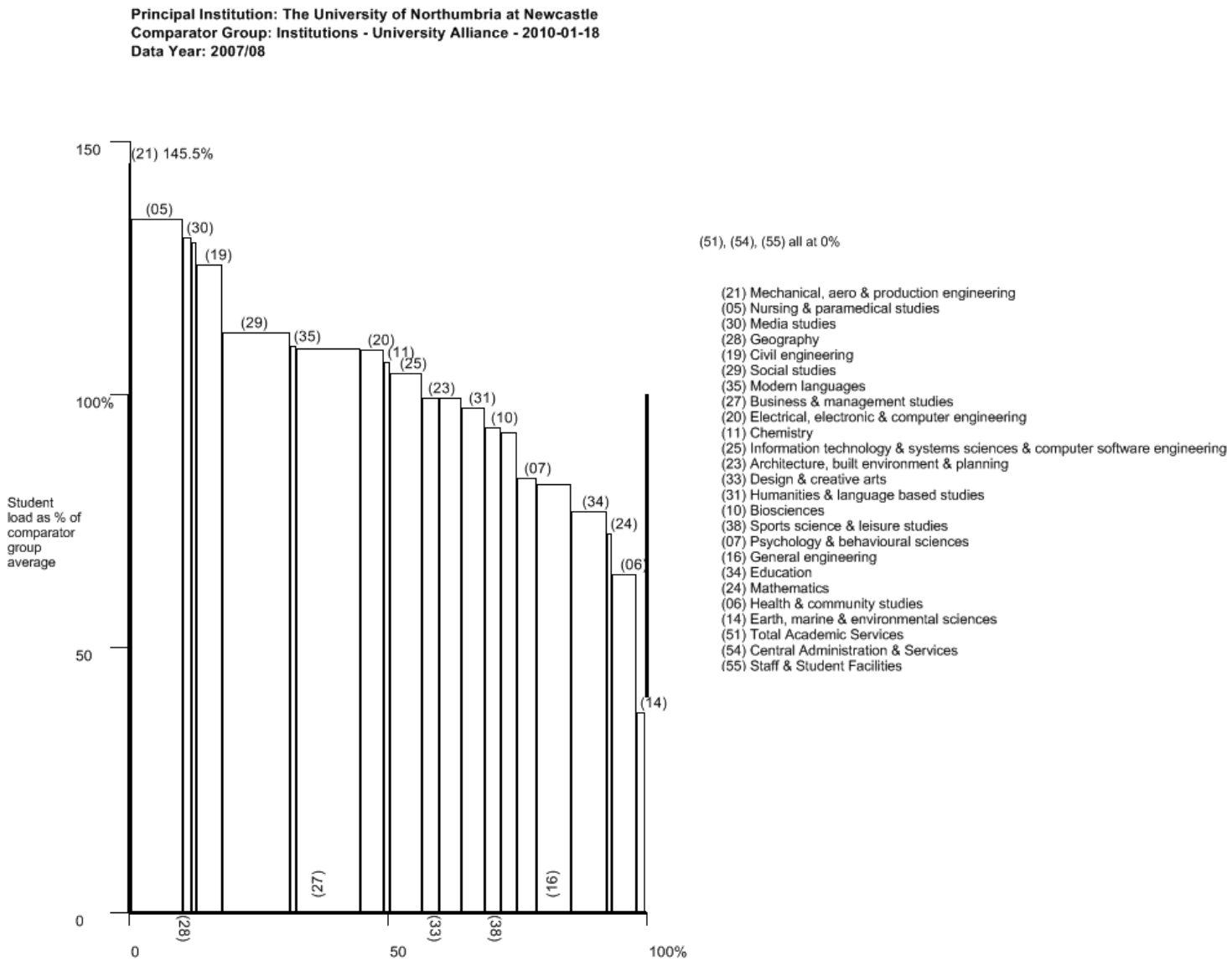


Figure 3: A Taylor cube produced by HEIDI. This is the current method used by Northumbria University to represent and extract data about the University

However, whilst graphs can be useful, they pose a problem in that they are unable to visualise data with many dimensions, such as the amount of data or dimensions that can be derived from the data sets in the National Student Survey (NSS). Therefore the task of effectively using the data and using the

"right" bits of data is becoming more and more important for organisations such as Northumbria University and they need a more effective way to monitor their performance, make more accurate predictions and identify where they could improve.

Additionally, one further problem arises when attempting to evaluate how effective a particular method is at visualising the data. Finding benchmarks, to evaluate the work against will form a major part of the evaluation process and will require considerable and in-depth research to determine whether the visualisation tool is providing is "better" results than previous incarnations of tools, and techniques currently in use by the client. Furthermore, trials with the client could be used to gauge opinion and to evaluate whether what has been built is actually what was required.

The problem that we look to solve here is representing more effectively and visually the data used by Northumbria University to monitor their own performance against other universities, and how well they are doing compared to their "aspirational" institutes, i.e. those they look up to and strive to be like.

Using data derived from data cubes developed from statistics taken from the National Student Survey (NSS) this research project aims to visualise and improve the current techniques and knowledge the university can currently gain from this data by applying visualisation techniques such as Targeted-Projection Pursuit (TPP) and such a technique may help reduce the risk of having to reduce the dimensions of the data (Faith, 2007) although research work will be conducted to determine which of the available techniques is the best option.

Why this project?

This project is worthwhile in that it looks to solve real-world issues, and looks to improve and build up on an area of intensive research where often there is much effort placed in to improving the way we can represent, use and manipulate data and where ways of evaluating solutions takes effort, thought and much testing. The research project is also worthwhile as it aims to answer a

big research problem that is it aims to produce a solution to an area where there has been little application to the problem, i.e. visualising OLAP cubes, something which is confirmed by Techapichetvanich and Datta (2005) in their paper *Interactive Visualization for OLAP*.

Attempting to contribute to this research area will be worthwhile for all involved and ultimately a tool that can actually be used by the client will be produced and will provide the University with a more effective method to visualise their data by applying data visualisation techniques and evaluate the effectiveness of the technique.

Currently, there are several tools available which have been identified that, with modification could be used or there may be scope to develop a tool from scratch. This will be investigated further during the analysis stage.

Project Aims

To apply a technique to visualise high-dimensional data sets taken from OLAP cubes and evaluate the effectiveness of the technique on the data derived from the National Student Survey and in doing so provide some improvement on the current techniques currently in use and allow for better visual display of the resulting data.

Project Objectives

1. To investigate existing techniques for visualising data from the National Student Survey
2. To investigate visualisation techniques for high-dimensional data
3. To investigate existing relevant software for visualising this particular type of data
4. To investigate appropriate development tools and techniques
5. To produce suitable designs for the proposed piece of software
6. To design an evaluation strategy
7. To implement the design and produce the software application
8. To test and evaluate the software application and its effectiveness on the data sets

9. To conduct an evaluation of the project process
10. To conduct an evaluation of the development methods and tools used during the development process
11. To produce a report on an ongoing basis

Research Areas

- Data Visualisation
- OLAP cubes
- Data exploration
- Database mining

Structure and Contents of Project Report

The report that will be produced will include several main chapters each with sub-sections. The main body of the report can be broken down into the following sections and as such each section will form the basis of the chapters;

- Introduction; this section of the report will introduce the project and will set out what work is planned throughout the project. [Objective 1]

- Literature Review; an overview of current research that is taking and what has been done. This will be split into several chapters. [Objective 1, 2, 3 and 4]

- Design; this section will include a discussion of all the design, and will layout the plans for the proposed solution. [Objective 5 and 6]

- Implementation and testing; this section deals with any implementation and testing work that has been carried out during the development of any software that is produced. [Objective 7 and 8]

- Evaluation; this particular section will provide a discussion of the product including an evaluation of it and the methods and

tools used to develop the software application. This section will also evaluate the research process and the procedures adopted during the project. [Objective 9 and 10]

- Conclusion, recommendations and further work; this section will discuss any conclusions that have been made from the research. It will discuss how future enhancements and how techniques used during the process could be implemented in any future work. [Objective 10]

Description of the expected product items

The following sets of deliverables are expected to be produced during this research project;

- A survey of the literature (literature review)
- A Software Requirements Specification
- An Analysis model
- A Project plan
- A Design model
- A comprehensive and exhaustive set of test plans and set of results
- An evaluation of the software application, techniques and tools used during the course of the project

Relationship to the Course

The project will incorporate many of the skills and knowledge learned through the study of the MSc Computer Science degree at Northumbria University. It will allow for the skills and knowledge gained through the study of modules such as Research Methods and Project Management to be applied to the dissertation and will allow knowledge gained in other modules, more particularly the more practical modules to be applied also. However, it is also thought that there will be the need to gain knowledge outside the scope of the course also, but that this work will compliment the skills already gained through the study of the MSc degree.

APPENDIX A

Discussion of any ethical issues

There may be the need to consider some ethical issues with the project. There may be the need to access sensitive data that may not be shared with others as well as accessing documents, computer systems and/or particular information about business processes that the client would not wish other individuals or organisations be aware of.

Whilst the data will not identify individuals or particular groups, there will be need for further discussion with the client to ensure that privacy is maintained and any ethical considerations are dealt with in an appropriate manner and that data is fully and properly protected.

Further information can be found on the copy of Ethics Form B in Appendix B.

Resources - statement of hardware / software required

It has already been agreed with the relevant parties that where required, access to specialist systems, university systems and/or specialist data such as NSS data will be given and the installation of any additional software will be made available and carried out by the appropriate university departments where and if required.

Regarding the production of the report and associated documentation, it is planned to use standard software including but not limited to Microsoft Office suite of products; including Microsoft Word & Microsoft Project and the use of a standard web browser is also expected and will be available at the university and other locations. These resources will be available both at the University. Additional hardware including printers will again be available at the University.

Sources of information, references and bibliography

Cheng, L. T., Kerr, B. (no date) *Project: Bloom.* Cambridge, MA: IBM Watson Research Center [Online]. Available at: http://domino.watson.ibm.com/cambridge/research.nsf/242252765710c19485256979004d289c/864fa36402234708852570f90079a47a?OpenDocument (Accessed: 22 March 2010).

APPENDIX A

Dubler, C. and Wilcox, C. (2002) *Just What Are Cubes Anyway? (A Painless Introduction to OLAP Technology).* Denver, CO: Microsoft Corporation [Online]. Available at: http://msdn.microsoft.com/en-us/library/aa140038(office.10).aspx (Accessed: 17 March 2010).

Faith, J. (2007)' Targeted Projection Pursuit for Interactive Exploration of High-Dimensional Data Sets' *11th International Conference Information Visualization* Zürich, Switzerland 2 - 6 July 2007. IEEE Computer Society

Friendly (2008) *Milestones in the history of thematic cartography, statistical graphics, and data visualization.* York, Ontario: York University [Online]. Available at: http://www.math.yorku.ca/SCS/Gallery/milestone/milestone.pdf (Accessed: 22 March 2010).

Ipsos MORI. (2010) *The National Student Survey.* [Online]. Available at: http://www.thestudentsurvey.com/faqs.asp (Accessed: 22 March 2010).

Mailvaganam, H. (2007) *Introduction to OLAP.* [Online]. Available at: http://www.dwreview.com/OLAP/Introduction_OLAP.html (Accessed: 16 March 2010).

Maniatis, A. S., Vassiliadis , P., Skiadopoulos, S. and Vassiliou, Y. (2003) ' Advanced Visualization for OLAP' *6th ACM international workshop on Data warehousing and OLAP* New Orleans, LA, USA 7 November 2003. New York, NY, USA: ACM.

Techapichetvanich, K. and Datta, A. (2005) 'Computational Science and Its Applications – ICCSA 2005', *Lecture Notes in Computer Science,* 3482/2005, 2 May 2005, pp. 206 - 214 *Computer Graphics and Rendering Workshop* [Online]. Available at: http://portal.acm.org/citation.cfm?id=956063 (Accessed: 20 March 2010).

Zaïane, O. R. (1999) *Department of Computing Science: CMPUT 690: Principales KDD (Glossary).* [Online]. Available at: http://webdocs.cs.ualberta.ca/~zaiane/courses/cmput690/glossary.html (Accessed: 23 March 2010).

Weskamp, M. (2004) *Newsmap.* [Online]. Available at: http://marumushi.com/projects/newsmap (Accessed: 23 March 2010).

Yang, K., Li, Y., Luo, Q., Sander, P. V. ands Shi, J. (2009) 'I3DC: Interactive Three-Dimensional Cube', *25th IEEE International Conference on Data Engineering, ICDE 2009*. Shanghai, China 29 March - 2 April 2009. Inst. of Elec. and Elec. Eng. Computer Society, pp. 1475-1478. [Online]. Available at: http://ieeexplore.ieee.org.wf2dnvr2.webfeat.org/stamp/stamp.jsp?tp=&arnumber=4812551 (Accessed: 17 March 2010).

APPENDIX A

ACM Special Interest Groups: SIGGRAPH: ACM Special Interest Group on Computer Graphics and Interactive Techniques

Higher Education Information Database for Institutions (HEIDI) available at: www.heidi.ac.uk

Data, and OLAP cubes will be supplied by the University's Corporate Planning Department.

Appendices

Appendix B – Ethics Form B

Appendix C – Gantt Chart

APPENDIX B

APPENDIX C

APPENDIX D

Question Number	Scale	Question
Q1		Staff are good at explaining things.
Q2		Staff have made the subject interesting.
Q3		Staff are enthusiastic about what they are teaching.
Q4	The teaching on my course	The course is intellectually stimulating.
Q5		The criteria used in marking have been clear in advance.
Q6		Assessment arrangements and marking have been fair.
Q7		Feedback on my work has been prompt.
Q8		I have received detailed comments on my work.
Q9	Assessment and feedback	Feedback on my work has helped me clarify things I did not understand.
Q10		I have received sufficient advice and support with my studies.
Q11	Academic support	I have been able to contact staff when I needed to.
Q12		Good advice was available when I needed to make study choices.
Q13		The timetable works efficiently as far as my activities are concerned.
Q14		Any changes in the course or teaching have been communicated effectively.
Q15	Organisation and management	The course is well organised and is running smoothly.
Q16		The library resources and services are good enough for my needs.
Q17		I have been able to access general IT resources when I needed to.
Q18	Learning resources	I have been able to access specialised equipment, facilities or room when I needed to.
Q19		The course has helped me present myself with confidence.
Q20		My communication skills have improved.
Q21	Personal development	As a result of the course, I feel confident in tackling unfamiliar problems.
Q22	'Overall, I am satisfied with the quality of the course'	Overall, I am satisfied with the quality of the course.

APPENDIX E

```
@relation UNNNSSDataByJACSSubjects

@attribute Year numeric
@attribute Subject {'L102 Subjects allied to Medicine','L103 Biological Sciences','L106 Physical
Sciences','L108 Computer Science','L110 Architecture Building and Planning','L111 Social studies','L112
Law','L113 Business and Administrative studies','L114 Mass Communications and Documentation','L115
Languages','L116 Historical and Philosophical studies','L117 Creative Arts and Design','L107 Mathematical
Sciences','L109 Engineering and Technology','L120 Initial Teacher Training','L121 Geographical Studies'}
@attribute 'Average Q1' numeric
@attribute 'Average Q2' numeric
@attribute 'Average Q3' numeric
@attribute 'Average Q4' numeric
@attribute 'Average Q5' numeric
@attribute 'Average Q6' numeric
@attribute 'Average Q7' numeric
@attribute 'Average Q8' numeric
@attribute 'Average Q9' numeric
@attribute 'Average Q10' numeric
@attribute 'Average Q11' numeric
@attribute 'Average Q12' numeric
@attribute 'Average Q13' numeric
@attribute 'Average Q14' numeric
@attribute 'Average Q15' numeric
@attribute 'Average Q16' numeric
@attribute 'Average Q17' numeric
@attribute 'Average Q18' numeric
@attribute 'Average Q19' numeric
@attribute 'Average Q20' numeric
@attribute 'Average Q21' numeric
@attribute 'Average Q22' numeric

@data
2007,'L102 Subjects allied to Medicine',
3.939394,3.77,4.09,3.828283,3.937081,3.71,3.58,3.64,3.434343,3.710679,3.811881,3.664125,3.336634,3.28,3.25,3
.823087,3.996894,3.600686,3.89617,4.121212,3.906239,3.73
```

APPENDIX E

```
2007,'L103 Biological Sciences',
4.08,3.656566,4.032925,4.03,3.717172,3.68,2.742574,2.87,2.79348,3.616162,4,3.673077,3.65,3.783673,3.73,4.09,
4.207921,3.918288,3.64,3.87,3.74,3.989899

2007,'L106 Physical Sciences' ,
4.11,4.02,4.34,4.18,3.96,3.96,3.080808,3.44,3.274638,3.93,4.242424,3.853585,3.806122,3.85,4.01,4.03,4.212121
,3.914717,4.050505,4.25,4.15,4.222222

2007,'L108 Computer Science',
4,3.74,3.99,3.93,3.94,4.04,3.168317,3.465347,3.574257,3.89,4.178218,3.835112,3.74,3.623762,3.700137,4.356699
,4.53,4.068833,3.970297,4.17,3.96,4.07

2007,'L110 Architecture Building and Planning',
3.910891,3.762376,4.009901,3.891089,3.643564,3.585859,3.04,3.58,3.401948,3.77602,3.994898,3.694286,3.825765,
3.318367,3.584158,4.260924,4.28,3.871858,3.907204,4.014796,3.935842,3.990099

2007,'L111 Social studies',
3.910891,3.676768,3.990099,3.816733,3.55,3.761474,3.149402,3.257426,3.10101,3.756972,4.065737,3.534661,3.788
955,3.505936,3.517928,3.828283,4.22,3.675492,3.90636,4.165339,3.926295,3.8

2007,'L112 Law',
4.12,3.76,4.01,4.363636,3.878788,4.049072,2.861443,3.118763,3.059975,3.68,4.118351,3.870928,3.79,3.777778,3.
878788,4.191919,4.48,4.117258,4.139525,4.31,4.21,4.18

2007,'L113 Business and Administrative studies',
3.881188,3.61,3.848485,3.765625,3.760379,3.548545,3.019894,3.12024,3.049471,3.578466,3.89863,3.494511,3.7326
73,3.397884,3.458783,4.16,4.242424,3.751618,3.85913,4.057196,3.846218,3.831683

2007,'L114 Mass Communications and Documentation',
3.79,3.57,3.811881,3.55,3.656566,3.667368,2.81,3.111111,3.117771,3.5,3.64,3.431826,3.580622,3.227965,3.10101
,4.09901,4.25,3.638349,3.871287,3.91,3.75,3.56

2007,'L115 Languages',
4.07,3.97,4.14,4.03,3.75,3.81,3.151515,3.490633,3.346847,3.95,4.079208,3.569135,3.841584,3.382694,3.514851,4
.060606,4.129279,3.85363,4.06036,4.19,3.990099,4.089109
```

APPENDIX E

```
2007,'L116 Historical and Philosophical studies',
3.84,3.777778,4.06,3.950495,3.555556,3.82,3.25,3.75,3.56,3.828283,4.09,3.434343,4.07,3.7,3.465347,3.43,4.424
242,3.594845,3.738281,4.013906,3.911483,3.9

2007,'L117 Creative Arts and Design',
3.633663,3.73,3.778991,3.704132,3.47,3.282828,2.871287,3.32,3.2,3.343434,3.383838,3.400682,3.694239,3.140826
,2.75,4.105432,4.212121,3.533333,3.8,3.93,3.712871,3.39

2008,'L102 Subjects allied to Medicine',
4.2,4.02,4.3,4.212121,4.178218,4,4.04,3.920792,3.663366,4.069307,4.21,3.995714,3.86,3.43,3.57,4.353535,4.252
525,4.047619,4.24,4.454545,4.24,4.151515

2008,'L103 Biological Sciences',
4.161616,3.919192,4.207921,4.131313,3.838384,3.821782,3.388717,3.515152,3.454545,3.92,4.29703,3.990066,4.16,
4.065789,4.08,4.444444,4.505033,4.194079,3.986743,4.137622,3.990099,4.21

2008,'L106 Physical Sciences',
4.287129,4.059406,4.28,4.232323,3.939394,4,3.292929,3.53,3.732673,3.950495,4.35,3.834949,3.772277,3.969981,4
.199235,4.36,4.27,4.037684,4.12,4.226947,4.078526,4.3

2008,'L107 Mathematical
Sciences',4.08,3.88,4.48727,4.25,4.07,3.762376,3.98,3.95,3.465347,4.14,4.24,4.41,3.67,3.85,4.16,4.376238,4.5
54455,4.224348,3.95,4.15,3.97,4.306931

2008,'L108 Computer
Science',4.2,3.910891,4.27,3.89,4.060606,3.979798,3.53,3.66,3.73,4.21,4.31,4.069307,4.07,4.019802,4.05,4.51,
4.56,4.333537,4.3,4.56,4.34,4.24

2008,'L109 Engineering and
Technology',4.01,3.878788,3.99,3.99,3.76,3.76,3.247525,3.373737,3.51037,3.84,3.993827,3.932407,4.08,3.685741
,3.673267,4.373737,4.480802,3.962223,3.979798,4.080808,4.03,4.01

2008,'L110 Architecture Building and
Planning',4.08,4.08,4.37,4.178218,3.9035,3.871287,3.711998,4.022643,3.782178,4,4.28,4.126596,4.24,4,3.930693
,4.464646,4.454545,4.222286,4.080808,4.2,4.14,4.22
```

APPENDIX E

```
2008,'L111 Social
studies',4.05,3.72,3.9,4.04,3.878788,3.969697,3.306931,3.53,3.36,3.88,4.252525,4.03,3.643564,3.25,3.564356,4
.59596,4.59,4.121285,3.95,4.02,3.98,3.9

2008,'L112 Law',
4.118812,3.940594,4.20202,4.117308,3.89,3.809949,3.610939,3.506322,3.282142,3.954936,4.216287,3.857455,3.917
788,3.872387,3.806084,4.10101,4.256971,3.837219,3.944884,4.127284,4.007648,4.08

2008,'L113 Business and Administrative studies',
4.09901,3.8,4.09,3.97,4.16,4.01,3.705333,3.618491,3.473581,3.884649,4.177129,3.83302,3.97,3.827981,3.927737,
4.42,4.46,3.988467,4.07,4.17,4.090909,4.181818

2008,'L114 Mass Communications and Documentation',
4,3.989899,4.148515,3.8,3.909091,3.979798,3.747475,4.086905,3.76,3.87,4.089109,3.77,4.057262,3.543452,3.54,4
.264762,4.333333,3.872857,4.108911,4.2,4.090909,3.96

2008,'L115 Languages',
4.26,4.13,4.35,4.292929,4.163536,4.138614,3.68,3.915755,3.863839,4.18,4.474747,3.970297,4.17,3.924107,3.9306
93,4.29,4.412155,4.025635,4.113812,4.3,4.191919,4.178218

2008,'L116 Historical and Philosophical studies',
4.18,3.998523,4.15,4.191919,3.818182,4.20202,3.693069,3.89,3.653465,3.90099,4.181818,3.94,4.127045,4.008409,
4.09,3.808081,4.26,3.699439,3.831683,3.96,3.86,4.20202

2008,'L117 Creative Arts and Design',
4.030303,4.06,4.262626,4.010101,3.8,3.823925,3.485148,3.848485,3.71,3.82,3.888791,3.846563,3.787879,3.582777
,3.25,4.165553,4.188752,3.58862,3.99,4.191919,4.00737,3.93

2008,'L120 Initial Teacher Training',
4.151515,4.14,4.57,4.34,4.34,3.930693,4.14,4.131313,3.747475,4.16,4.27,4.140166,3.83,3.88,4.04,4.17,4.181818
,3.903778,4.28,4.52,4.31,4.41

2008,'L121 Geographical Studies',
4.141414,4.21,4.44,4.22,4.03,4.05,3.666667,3.93,3.9,4.12,4.30303,4.04,3.86,3.89,3.930693,4.333333,4.09,3.570
013,4.15,4.363583,4.18,4.33
```

APPENDIX E

```
2009,'L102 Subjects allied to Medicine',
4.080808,3.96512,4.09486,4.15,4.24,4.025187,3.87464,3.909091,3.643564,4.01,4.080808,3.956781,3.47,3.298598,3
.37,4.41,4.36,4.089533,4.252525,4.505051,4.3,4.06

2009,'L103 Biological Sciences',
3.960396,3.62,4.03,3.89,3.6,3.560462,3.111111,3.388641,3.117992,3.66,4.04,3.744111,4.016446,3.953101,3.97,4.
24,4.25,3.995122,3.87,3.98,3.89899,3.969697

2009,'L106 Physical Sciences',
4.06,4,4.14,4.11,3.707071,3.75282,3.12,3.316832,3.325823,3.808081,4.13,3.859095,3.72,3.93,3.95,4.37,4.3,4.11
0066,4,4.04,4.03,4.131313

2009,'L107 Mathematical Sciences',
4.26,3.94,4.46,4.54,4,4.343434,3.84,4.27,3.69697,4.191919,4.69697,4.29,4.53,4.262485,4.48,4.323232,4.56,4.27
1977,4.19,4.31,4.28,4.47

2009,'L108 Computer Science',
4.04,3.727273,3.91,3.747475,3.919192,3.62,3.363636,3.356436,3.26,3.719718,3.989899,3.49,3.63,3.680282,3.4343
43,4.202817,4.35,4.043944,3.808081,4.075391,3.896159,3.74

2009,'L109 Engineering and Technology',
3.81,3.57,3.727273,3.742574,3.585859,3.666667,3.316832,3.36,3.452037,3.75,3.846878,3.657705,3.87,3.527104,3.
467432,4.09,4.213333,3.898907,3.745902,3.868687,3.8,3.7

2009,'L110 Architecture Building and Planning',
3.96,3.87,4.20202,3.99,3.6,3.66,3.168903,3.645806,3.46,3.871287,4.06,3.903143,3.888889,3.67,3.67,4.415842,4.
21,3.908129,4,4.148515,4,4.009901

2009,'L111 Social studies'
,4.171717,4.029703,4.29,4.06,3.91,3.927842,3.90099,3.79,3.57,4,4.237123,3.87321,3.96,4.06007,4.1,4.138614,4.
264107,3.874664,4.03,4.2,4.08,4.14

2009,'L112 Law',
4.069307,3.86,4.05,4.22,3.6,3.754899,2.970201,3.009933,3.160521,3.787879,4.12,3.74,3.88,3.72449,3.97,4.05,4.
21,4.048456,4.168317,4.306931,4.262626,4.26
```

APPENDIX E

```
2009,'L113 Business and Administrative studies',
3.871287,3.572669,3.86514,3.658156,3.64,3.309134,3.07,3.23,3.04,3.558436,3.904916,3.496334,3.807737,3.475978
,3.356436,4.336634,4.352402,3.971652,3.857598,4.09,3.924804,3.69
2009,'L114 Mass Communications and
Documentation',4.02,3.88,4.11,3.7,3.77,3.79,3.52,3.66,3.43,3.61,3.74,3.445263,3.492807,3.377716,3.277817,4.1
00351,4.129825,3.46,3.913684,3.923509,3.820175,3.65
2009,'L115 Languages',
4.150968,3.979798,4.191919,4.18,4.08,4.020202,3.88,3.909091,3.727273,4.026235,4.37,3.914026,4.02,4.016623,3.
95,4.16,4.192208,3.888423,4.04,4.194286,4.088377,4.18

2009,'L116 Historical and Philosophical studies',
4.06,4.009901,4.207921,4.16,3.979798,4.12,3.831683,3.891089,3.63,3.891089,4.27,3.66137,4.158416,3.789041,4,3
.58,4.183562,3.561644,3.861386,3.95,3.898535,4.12

2009,'L117 Creative Arts and Design',
3.811881,3.888889,3.9941,3.821782,3.64,3.55,3.49,3.59,3.545455,3.79798,3.78,3.681173,3.76,3.444444,3.23,4.04
3805,3.984315,3.561593,3.837522,4.046903,3.904631,3.623762

2009,'L120 Initial Teacher Training',
4.217822,4.08,4.57,4.111111,4.23,3.83,3.69,3.841584,3.57,4.21,4.42,4.305333,3.83,3.868687,4.050505,4.05,4.07
,3.930245,4.384444,4.433889,4.414111,4.33

2009,'L121 Geographical Studies',
4.242424,4.049505,4.5,4.08,4.1,4.01,3.434343,3.86,3.717172,4.35,4.425743,4.2,3.949495,4,4.23,4.343434,4.38,4
.256727,4.158416,4.252525,4.151515,4.28

2010,'L102 Subjects allied to Medicine',
4.067511,3.966245,4.182203,4.180672,3.932489,3.666667,3.835443,3.281481,3.20155,3.447552,3.355769,3.378378,3
.167969,3.186992,3.149425,3.36036,3.35119,3.376206,3.303406,3.224932,3.271341,3.384868

2010,'L103 Biological Sciences',
4.021978,3.736264,4.114286,3.931868,3.830769,3.545055,3.406593,3.191033,3.164969,3.326415,3.288591,3.289963,
3.284497,3.325132,3.363478,3.313178,3.309791,3.371479,3.342857,3.320273,3.358792,3.384886
```

APPENDIX E

```
2010,'L106 Physical Sciences',
4.114943,3.977011,4.218391,4.034483,3.804598,3.863636,3.413793,3.23,3.268817,3.37963,3.301587,3.336283,3.31,
3.293103,3.388889,3.34188,3.295652,3.227273,3.275862,3.288136,3.274336,3.324786

2010,'L107 Mathematical Sciences',
4.285714,4.107143,4.448276,4.344828,4.357143,4.178571,4.206897,3.25,3.405405,3.302326,3.234043,3.2,3.209302,
3.2,3.333333,3.238095,3.410256,3.514286,3.35,3.25641,3.333333,3.217391

2010,'L108 Computer Science',
3.910891,3.762376,4.009901,3.891089,3.643564,3.585859,3.04,3.58,3.399605,3.78,4,3.697021,3.83,3.32,3.584158,
4.253686,4.28,3.919091,3.911765,4.02,3.940594,3.990099

2010,'L109 Engineering and Technology',
3.853741,3.659864,3.825939,3.863946,3.765306,3.638225,3.506803,3.204204,3.222571,3.390029,3.327177,3.338109,
3.253886,3.245014,3.286567,3.206651,3.270732,3.281843,3.34903,3.283887,3.352151,3.362606

2010,'L110 Architecture Building and Planning',
4.114943,3.977011,4.218391,4.058824,3.804598,3.883721,3.418605,3.23,3.268817,3.383178,3.301587,3.351852,3.31
3131,3.295652,3.388889,3.34188,3.298246,3.247525,3.280702,3.290598,3.279279,3.324786

2010,'L111 Social studies',
4.12,3.76,4.01,4.356537,3.878788,4.054536,2.860362,3.12,3.060606,3.678243,4.13,3.868747,3.787959,3.775748,3.
878788,4.191919,4.476176,4.190342,4.14552,4.306615,4.203763,4.18

2010,'L112 Law',
3.61,3.848485,3.767677,3.762376,3.55,3.02,3.121212,3.05,3.58,3.90099,3.515152,3.732673,3.4,3.46,4.16,4.24242
4,3.867518,3.861386,4.06,3.842363,3.831683,?

2010,'L113 Business and Administrative studies',
3.79,3.57,3.811881,3.55,3.656566,3.673267,2.810394,3.110878,3.118568,3.5,3.633416,3.42429,3.585859,3.23,3.10
101,4.096752,4.247407,3.665712,3.867715,3.908112,3.746894,3.56

2010,'L114 Mass Communications and Documentation',
4.07,3.97,4.14,4.03,3.75,3.81,3.151515,3.49505,3.35,3.95,4.079208,3.574257,3.841584,3.383306,3.514851,4.0606
06,4.141481,3.976057,4.07,4.19,3.982837,4.089109
```

APPENDIX E

```
2010,'L115 Languages',
3.84,3.777778,4.06,3.950495,3.555556,3.82,3.25,3.746269,3.557214,3.828283,4.09,3.432161,4.07,3.689655,3.4653
47,3.43,4.41,3.60241,3.742574,4.024876,3.931373,3.9

2010,'L116 Historical and Philosophical studies',
3.633663,3.73,3.782178,3.71,3.47,3.282828,2.871287,3.32,3.2,3.343434,3.383838,3.40404,3.7,3.141414,2.75,4.11
,4.212121,3.465517,3.8,3.905844,3.712871,3.39

2010,'L117 Creative Arts and Design',
4.285714,4.107143,4.448276,4.344828,4.357143,4.178571,4.201222,3.25,3.404656,3.302326,3.234043,3.19879,3.209
302,3.2,3.333333,3.237708,3.40882,3.507336,3.35,3.255513,3.331586,3.217391

2010,'L120 Initial Teacher Training',
4.32,4.36,4.65,4.46,4.08,3.91,4.17,3.297297,3.258929,3.304348,3.210191,3.253333,3.235294,3.323741,3.326241,3
.214286,3.253247,3.310606,3.253247,3.256579,3.293333,3.181818

2010,'L121 Geographical Studies',
4.166667,4.055556,4.555556,4.222222,4,4.111111,3.638889,3.369565,3.318182,3.333333,3.218182,3.26087,3.183673
,3.254902,3.306122,3.236364,3.25,3.22449,3.269231,3.15,3.283019,3.226415
```

APPENDIX F

Glossary of terms

ARFF	Attribute-Related File Format
CPD	Corporate Planning Department
HECFE	Higher Education Funding Council for England
HEI	Higher Education Institution
HEIDI	Higher Education Information Database for Institutions
JACS	Joint Academic Coding System
MDS	Multidimensional Scaling
NSS	National Student Survey
PCA	Principal Component Analysis
TPP	Targeted Projection Pursuit