Working Project Title – Data Mining and Data Warehousing within the British Transport Police.

Introduction

The Central Justice Unit (CJU) in London collates and records a large quantity of data that I feel contains untapped knowledge and information.
The CJU are also under-going a major change to becoming a member of the Institution for Customer Service.

With these statements in mind I am developing a project whereby I am able to utilise the knowledge gained throughout my Data Mining Module in the first semester of year 3, find knowledge in the mass of data collected as well as to demonstrate data mining techniques in a live environment. The main aim for the use of the data kept by the CJU is to improve efficiency of the Performance Units tasks and knowledge of their customer base.

As part of this project I will also be researching the effectiveness of Data Warehousing and how it is used with a Police environment.

Daniel Stamate to whom I will have regular meetings with to discuss progress and possible techniques to complete an effective analysis will supervise this project.

Aims and Objectives

The aim of my project is to investigate, research and report Data Warehousing within a police environment and also utilise Data Mining to improve knowledge for the CJU.

The three areas I will be investigating within the CJU are:

- The inter-department relationship between the Performance Unit and the Witness Care Unit's (WCU's).
- The efficiency of obtaining court results by the Performance Unit
- Predict workload for the Performance Unit to ensure a resilient service is provided to the WCU's

My first objective is to collate the data required to perform the analysis. To cluster and classify customer needs and requirements I will be utilising an on-line survey to collect quantitative data to use for analysis as well as collecting qualitative data to provide additional information on the customers requirements of the WCU's.

To measure the efficiency of the performance unit I have made a record of each result obtained from court (downloaded from the court system Libra) since November 2011.

To assist with workload prediction I will also be using an existing record of the number of result requests that are printed each day since November 2009.

Taking into consideration the areas of investigation and data the questions I set out to answer are:

- What are the expectations of the Performance Unit from the WCU's?
- Are the Performance Unit downloading/obtaining court results at the optimum time?
- What are the busiest days/months/quarter for the Performance Unit?

My deliverables for my project will come in the form of a recommendations report, which will outline the results of my analysis and include suggestions on how the Performance Unit can make its processes more efficient as we as more appropriate for its customer base. I will also be developing programmes in Java to perform data mining on the obtained data.

Methods

I have three areas to analyse and after discussion with my supervisor it was a concern to be able to programme 3 data mining tools in time for the completion of the project. With this in mind I have decided to develop a Java written programme to analyse the time series/workload aspect of my project. I will be conducting further research in the options I have to fulfil this requirement, however a technique called windowing looks to be the most appropriate. I will be using Eclipse to write my code in Java.

For the other two areas of analysis I will look to use the open source software WEKA. I will use this package to develop multiple models for each data set, analyse their effectiveness and recommend the best results I receive to the management team.

The rational behind these decisions is to ensure that I demonstrate my understanding and coding abilities but also have a complete project. With the time constraints on this project and the inclusion of a two-week placement within the Development Team within the British Transport Police researching the use of Data Warehouses I have decided to us a combination of my own coding and a library to fulfil the requirements of the project.

I chose these methods, as I feel confident writing code in JAVA and have experience of WEKA through my third year module on Data Mining. I felt It was important to be confident with the software and coding languages I am going to use as I wanted to focus on the other areas for my project, including Data Mining and Data Warehousing.

Project Plan

To plan my project I have produced a Gantt chart to illustrate the tasks required to complete, including timescales. The overall plan to complete the project successfully is to begin coding as soon as possible, tackle each area of the investigation and continue to research simultaneously. The Gantt chart can be found alongside this report.

Progress to Date

This project involves a wide range of elements that require organising before programs and analysis can begin. My progress to date includes the following:

- Area of Analysis research
- Data Collection
- Data Mining Techniques research
- Data Warehousing and Data Cube Research
- Software selection and setup
- Project Plan

One of my first tasks for my project was to research the processes and tasks that were of interest to the Management Team of the CJU with regards to the inter-team customer service analysis and also the internal processes conducted by the Performance Unit that I and the Manager felt could be improved.

To do this I have held a meeting with the Data Manager of the Performance Unit and discussed a number of the teams' daily activities, how they are performed and when. From these meetings it came to light that although the team deliver a high level of service there could be room for improvement if we research and analyse the time taken between a court hearing being completed and the team downloading the result for that hearing from the court systems. The second area to research and analyse that derived from these meetings would be the capability of predicting workload throughout a working year, thus having more knowledge of staffing levels and job role delegation.

I have also discussed my project and plan during the bi-weekly managers meetings. In this forum we have discussed the areas in which an analysis would be necessary. I am in the process of writing a survey that will be completed by all the Witness and Case Officers (WACO). This survey will build my dataset and will be created from a set of questions developed by me and the managers of the WCU. They will be quantitative to enable to be process the outcomes effectively. This survey will look to bridging the gap between the expectations of the Performance Unit from the WCU and the service to be provided by the Performance unit to the WCU.

To complete my analysis on workload prediction I will be using historical data that is collated by the members of the Performance Unit. This data consists of the number of every result request that has been printed for each day since November 2009. This spreadsheet has been simplified and has been broken

down to the amount of London, county based and crown court hearing requests for each day. This data set is on going and will be updated until February 11th (Data Starts on November 11th thus chosen to create a full month).

To perform the analysis for classifying the optimum time to download court results from the court system I have collated data since November 2012 that defines the court hearing date, the court, the date the result was printed and the date and time of the previous check made by the team to download the result.

To date my research has been mostly formed of reading Data Mining: Concepts and Techniques by Jiawei Han and Internet research of terms. I have been collating and note taking of possible techniques that could be suitable for the data sets I have collected. I will be researching throughout the first half of my project to ensure that I am using suitable data mining and if necessary data warehousing (or data mart) techniques to produce the best results of my analysis. Thus far the areas of my research are:

- Time-series and Time-series forecasting
- Data Marts
- Trend Analysis in Time Series Data
- Data and Prediction Cubes
- Regression – including Linear and Multiple Linear regression and Method of least squares
- Windowing

Planned Work

The project is well underway. The data necessary for the project has been collected and the next tasks regarding data is to complete research on surveys and their structure, then using this research construct a survey on areas the department deem of interest. This will give the WACO's a significant period to answer the questionnaires and time to collate the data.

I will be looking to start constructing my Java program for the workload analysis consecutively to constructing the survey. I have already set-up my working environment on eclipse with a version control repository created.

These two tasks are the main features of the project so I felt it necessary to begin these as early as possible. With my next tasks being the main body of the project I will be looking to have a permissions letter written up by management British Transport Police, and look to have ethics forms filled in as soon as possible.

Reference List

Jiawei Han and Micheline Kamber and Jian Pei – "Data Mining Concepts and Techniques – Third Edition", 2011.

T. Taylor – "Bsc and Diploma in Computing and related subjects – Project" University of London guidance paper.