**Mining HKUCS Graduate Student Data****: Extraction, Analysis and Prediction**

**Project Plan (Draft)**

**Wu You (Johnson)**

**ABSTRACT**

This Project plan provides a comprehensive summarization of our final year project, Mining HKUCS Graduate Student Data: Extraction, Analysis and Prediction. HKUCS stands for the Computer Science Department of the University of Hong Kong.

The major objective of this project is to deliver a robust educational data mining web application, which allows users to do data extraction, analysis and prediction. The application can be used to mine the HKUCS graduate student data and greatly help the department ameliorate its admission strategies and facilitate the admission process.

This project plan gives the background information and related studies of this project. It also identifies the objectives and scope of this project and describes the approach and methodology used in this project. Information about project management and schedule is also included.

**INTRODUCTION**

Educational data mining (EDM) is a growing research area that aims at discovering useful information and studying educational problems by exploring data originating from educational context. [1] It is generally acknowledged that substantial valuable information, which could be vital to many educational decisions, is embedded in the educational data. Unfortunately, EDM is still not playing a part in the educational decision-making process in HKUCS. Each year, HKUCS attracts hundreds of applications from students all over the world for doing a graduate study. From these applications, numerous data have been collected, including student profiles, academic performance, research interests, scholarships, teachers’ comments and admission decisions. However, these data have not been well use of since there lacks an EDM tool in the current admission system to mine these data and probe for useful underlying information. If an EDM application can be built, the administrators and teachers of HKUCS can visualize and analyze the candidates’ data, and thus discover significant patterns or trends that can provide insights into the admission strategies as well as admission decision-making.

Therefore, the objective of this project is to develop an educational data mining web application, which can use graduate students’ data to conduct data extraction, visualization and analysis. Moreover, based on the analysis, the application can rank applicants by their performance, filter their applications and predict admission results. With these functionalities, the EDM application can greatly facilitate the admission process of HKUCS and help teachers make admission decisions.

This paper proceeds as follows. It first considers the existing researches in the field of EDM and explicates how their achievements can be incorporated into our project. Then it identifies the objectives and the scope of this project and describes the approach and methodology used in this project. It also lists the potential challenges in this project and proposes corresponding countermeasures. In the end, it provides information about the project management and schedule.

**RELATED STUDIES**

Currently there is a growing interest in educational data mining. Various studies have been carried out in this field.

Mashat et al. designed a model to do association rule mining on the university admission data and extract relations between different attributes like GPA and courses already taken by the students. [2] This model provides insights into how to compare the importance of different assessment criteria during the admission process. The use of association rule mining is adaptable in our project.

Feng et al. established a university admissions decision-making model by utilizing the Self Organizing Map (SOM) neural network, cluster analysis, association rule and Fayyad data mining model. [3] Their model astutely takes geographical data into consideration and develops a new reference for admission scheme and propaganda. [3] Their experiment on applying SOM to analyze geographical data draws lessons on how to make use of the origins of the applicants in our project.

Sharma and Bvaghela presented a new algorithm, which makes use of Global Rule Binary Search Tree and distributed data mining techniques to predict the result of students' admission to college. [4] This algorithm can be of enlightening significance to the development of the prediction function of our application.

**OBJECTIVES**

This project aims to develop an advanced data manipulation application, which enables the administrators or professors of HKUCS to interact with the graduate student data, do data mining and analysis.

The application is vital to the HKUCS graduate student admission. It can significantly facilitate the admission process and help assessors make better admission decisions by providing a list of functionalities, including data visualization, pattern analysis, correlation analysis, performance analysis, student comparison, performance ranking, smart filtering, intelligent matching and outcome prediction.

The application is scalable and can be later adapted to be used by the university to admit undergraduate students, or used by companies for recruitment purposes.

**SCOPE**

In this project, an educational data mining web application will be developed. Broadly speaking, the application has three layers, data extraction, data analysis and prediction.

In the data extraction layer, graduate student data will be extracted from the HKUCS database and presented to users in the form of statistical graphics, plots or information graphics. Moreover, some open-source data like university ranking will also be extracted and visualized. Users can interact with these visual representations of data and reason about the underlying information in these data.

In the data analysis layer, different analytic tools, such as pattern analysis, correlation analysis and performance analysis, will be developed. Along with these tools, a set of data mining algorithms and statistical models will be implemented. Furthermore, an elegant user interface will be developed so that users can easily and efficiently conduct these analyses.

In the prediction layer, the following five functionalities will be provided:

* Student comparison lets users to compare students’ performance based on the specific criteria listed by the user.
* Performance ranking provides the ranking of all the applicants by considering their performance in each criterion and the corresponding proportion assigned to each criterion by the user.
* Smart filtering helps filter students’ applications by different requirements and constraints.
* Intelligent matching provides an efficient tool to match applicants with professors based on both parties’ preferences.
* Outcome prediction predicts the admission result of each applicant and helps professors make admission decisions.

In this project, only the HKUCS graduate students’ data will be used for analysis and only those data mining techniques that have impacts on HKUCS admission will be examined and implemented. Other data mining techniques would not be considered.

**APPROACH AND METHODOLOGY**

For this project, the Agile Software Development approach is applied in order to ensure the data mining application can satisfy the needs of HKUCS in terms of the graduate student admission. During the whole project, we will closely work with the administrators and professors of HKUCS to understand their requirements and seek feedbacks for our designs and implementations.

For the data extraction part, we will first examine existing open-source data visualization tools like Charted and D3 and make use of the appropriate ones to represent and visualize the HKUCS graduate student data.

For the data analysis part, we will research existing data mining techniques and algorithms by reading relative publications and papers. We will also take relative online courses to better understand different techniques. Then, we will develop prototypes to inspect the performance of different data mining algorithms and identify those that meet our requirements. In the end, based on the selected algorithms, we will develop the final version of the data analysis tool with a nice interface.

For the prediction part, we will implement the five functionalities in the sequence of Smart Filtering, Student Comparison, Performance Ranking, Outcome Prediction, Intelligent Matching. In this way, later work can be build upon the previous one.

The implementation of this web application is divided into front end development and back end development. JavaScript, CSS and HTML will be used as the languages for front end development, which is responsible for visualizing data and interacting with users. Meanwhile, different open-source JavaScript libraries like d3.js will be used. Python will be used for back end development, which is responsible for interacting with database, implementing data mining algorithms and analytical logic. In particular, ScikitLearn, as a Python library that provides a number of functionalities of data mining and data analysis, [5] will be used in this project.

**CHALLENGES AND MITIGATION**

One of the potential challenges of this project is to correctly identify the appropriate data mining algorithms and models that fit our purpose. There are many existing data mining algorithms and using an inappropriate algorithm might result in a useless analysis. To mitigate the deficiencies brought by this challenge, we will first try to understand each algorithm and consult experts on the use of these different algorithms. Later, we will make protocols to examine the performance of each algorithm with our HKUCS graduate student data and make the final decision on using which algorithm.

Another challenge lies in the paucity of data. Now, only the data of the graduate students in the last three years have been collected. Therefore, only limited implications can be provided with insufficient data. To address this problem, we will try to find alternative data like the graduate student data from other departments under the Faculty of Engineering.

**DELIVERABLE**

The deliverable of this project is an educational data mining web application that connects to the HKUCS database and provides all the functionalities listed in the SCOPE session.

**EVALUATION**

The web application delivered by this project will be evaluated according to the performance of each of the three layers.

For the data visualization layer, we will invite administrators and professors of HKUCS to determine whether the application makes the data more understandable and usable.

For the data analysis layer, we will test whether the application can identify inconspicuous patterns, trends and correlations and whether these discoveries can help HKUCS develop admission strategies.

For the prediction layer, we will compare the rankings, predictions and matching results provided by the application with the actual rankings and admission results for the past three years and test its accuracy. If it achieves an 80% accuracy, this layer will be considered as completed.

**SCHEDULE**

|  |  |
| --- | --- |
| Sep | * Project plan * Project website |
| Oct | * A prototype with the following implementations:   + one data visualization tool   + one data analysis tool with a specific data mining algorithm |
| Nov - Dec | * A prototype with more data visualization tools and data analysis tools |
| Jan | * A web application with all the determined data visualization tools and data analysis tools * Interim report |
| Feb - March | * A web application with the following five functions implemented:   + Smart Filtering   + Student Comparison   + Performance Ranking   + Outcome Prediction   + Intelligent Matching * Integration into the existing HKUCS system |
| April | * Evaluation and testing of the web application * Final report * Project presentation |
| May | * Project exhibition |

**WORK DIVISION**

|  |  |
| --- | --- |
| Data visualization | Xu Fangyuan |
| Data analysis | Wu You |
| Prediction | Xu Fangyuan |

**CONCLUSION**

This project will develop a robust educational data mining web application for the administrators and teachers of HKUCS to analyze the graduate student data in order to more efficiently admission decisions and answer questions that can be crucial to the admission strategies of HKUCS. We believe this application will have a real impact on the admission process in the future and can be later adapted to help universities to admit undergraduate students, or help companies recruit new employees.

**REFERENCES**

1. Romero C, Ventura S. Educational Data Mining: A Review of the State of the Art. IEEE Trans. Syst., Man, Cybern. C. 2010; 40: 601-618.
2. Mashat AF, M.fouad M, Yu PS, Gharib TF. Discovery of Association Rules from University Admission System Data. IJMECS International Journal of Modern Education and Computer Science. 2013; 5:1-7.
3. Feng S, Zhou S, Liu Y. Research on Data Mining in University Admissions Decision-making. International Journal of Advancements in Computing Technology IJACT. 2011; 3: 176–186.
4. Bvaghela D, Sharma P. Students' Admission Prediction using GRBST with Distributed Data Mining. Communications on Applied Electronics CAE. 2015; 2: 15–19.
5. Nelli F. Machine Learning with scikit-learn. Python Data Analytics. 2015; : 237–264.