Abstract

This project plan provides a comprehensive overview of the final year project, Mining HKUCS Graduate Student Data: Extraction, Analysis and Prediction. HKUCS is the abbreviation of the Computer Science Department of the University of Hong Kong.

The main objective of the project is to create a web tool that uses data mining techniques to find the patterns related to admission decision and application trend in the data. To achieve the goal, the web tool will provide a set of functions that visualize application data, analyze the data and predict the application outcome. The details of the functionalities to be realized are included in the project plan. Based on the objectives, the project plan examines the related studies. It also considers some potential challenges and suggests possible solutions. To provide a more concrete idea of the implementation of the project, it discusses the methodology, deliverables and the distribution of work between teammates.

Contents

1 Introduction 2
2 Related Studies 2
3 Scope 2
4 Prerequisites 3
5 Methodology 3
6 Deliverable 4
7 Challenges and Mitigation 4
8 Schedule 5
9 Work Division 5
10 Conclusion 6
1 Introduction

Every year, the Computer Science department of the University of Hong Kong receives hundreds of applications for graduate programs, including Master of Philosophy and Doctor of Philosophy in Computer Science. The number of applications usually exceeds the capacity. Thus, the faculty has to identify the most suitable applicants and make admission decisions. There are mainly two challenges in the process. Firstly, it is not an easy task to evaluate a candidate given the complexity of the large range of factors considered. Parameters of different types, such as their GPA, their English skills and their research interests, correlate with each other and all contribute to the decision making. Apart from that, as an applicant usually makes multiple applications for different universities, there does exist the risk that the offer will be rejected by the applicant. In this case, the admission officer has to consider the tradeoff between admitting the best student and ensuring that the offer will be accepted.

The project will address the problem by making sense of applicants data with data mining techniques, which will provide useful insights, such as admission patterns and important attributes, for making the admission decisions. The goal of the project is to build a web tool that helps the users to extract useful data, analyze the patterns and predict the admission outcome with a set of functions. The web tool will be integrated into the existing web-based admission system.

2 Related Studies

Educational Data Mining has been an emerging discipline these years. It is concerned with data that comes from educational settings and uses data mining methods accordingly to discover useful information for decision making. In general, the methods most commonly used in the field include association rules, classification and recommender systems. [1] There are some existing projects that build data mining tools with these models and generate satisfying outputs.

Wakil et al. design a web recommender system that proposes suitable courses to students in private universities in Iraq. There are mainly three recommender approaches, Decision Tree, Neural Network and a fitness equation which calculates a weighted sum of some factors. The fitness equation is derived based on the previous enrollment data. The system outperforms existing recommendation system in terms of the comparison between the recommendation and actual enrollment. [2]

Mashat et al. develop a set of association rules from university admission system data. They derive a set of rules that appear frequently in the dataset using Apriori Algorithm. The rules include combinations of different parameters that are crucial for admission, such as a certain level of GPA and some related courses taken previously. The method is also adaptable in this project. [3]

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. [4] Their model astutely takes geographical data into consideration and develops a new reference for admission scheme and propaganda. 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.

Other methods such as linear and logistic regression will also be considered and studied in the project.

3 Scope

Given the complexity of the problem, there are different perspectives to understand the data. Thus, the scope of the project is divided into the following three layers.

The first layer is to extract data that are identified as useful based on the interviews with the users from the database and present them in an understandable way. The web tool will visualize the data in different graphical forms. Apart from that, users can interact with the visualization by actions such as mouseover to easily view further information. The forms and the interactive functions are to be determined based on the users needs.
With the data extracted, the web tool will provide users with different methods to analyze it and discover the patterns. It will implement a set of data mining algorithms, such as Decision Tree and Association Rules, on the student data. On top of that, it will create a user friendly interface for the users to easily conduct the analysis.

Based on the patterns found with data mining algorithms, the web tool will apply them to the incoming application data and present them from different perspectives to facilitate the admission decisions. The functions provided are listed as follows:

- **Smart filtering.** This function allows the users to filter the applicants based on one or more constraints, such as applicants research interest and GPA.

- **Intelligent matching.** This function suggests candidates to a specific professor based on their research interest and patterns found in the students previously admitted to this professor.

- **Outcome prediction.** This function is a classifier that can make predictions based on a numerical value generated by the tool. A ranked list of candidates will be suggested to the user as potential candidates, potential candidates for short-listed interview, potential candidates for a program and potential candidates who will accept the offer.

4 **Prerequisites**

1. **Scikitlearn** [5]: a python data mining library
2. **d3.js** [6]: a JavaScript library for visualizing data with HTML, SVG and CSS
3. **Database access**: data of the past three years’ graduate admission

5 **Methodology**

To start with, we will gather user requirements by interviewing the heads of graduate admission to better understand aspects that are crucial to admission. In the meantime, we will gather the existing admission data, which includes 30 attributes for each applicant. As they are in various forms, we will preprocess some of the attributes by converting them into numerical values by methods such as text mining. The access to the database is granted by our supervisor, Reynold Cheng.

Then we will focus on selecting data mining methods for the web tool and building the classifier. The data will be divided into training data and testing data. Both sets will be separated into 3 categories, applicants admitted who accepted the offer, applicants admitted who declined the offer and applicants not admitted. Based on the existing papers and works mentioned in Related Studies, we will make prototypes and apply data mining algorithms, such as Decision Tree, Association Rule and Logistics Regression on the three data sets respectively. Specifically, we will focus on the admission pattern of each professor to build the Intelligent Matching tool. We will then analyze the performance of each algorithm and determine the most suitable ones for the data. For the analysis part, we will present our models to the users to evaluate whether they meet the requirements. For the prediction part, we will run the tool on testing data and compare the prediction generated by the tool with the actual results and calculate the accuracy.

The implementation of the web tool will be divided into front-end and back-end development. We will use Python as the back-end language, which handles interaction with the database and implementation of the data mining tools and classifier. Specifically, we will use ScikitLearn, a Python library that contains different modules such as feature normalization, text mining and classification methods.[7] The main front-end functionalities are data visualization and interactive functions. We will use d3.js, a JavaScript library that combines powerful visualization components and a data-driven approach to DOM (the Document Object Model for HTML), which facilitates the implementation of interactive functions.
6 Deliverable

The deliverable of the project is a web tool integrated into the existing web-based admission system for the HKUCS graduate programs. The web tool communicates directly with the database and provides a user-friendly interface with a set of functionalities that help make sense of the data. Progress can be checked on the project website: http://i.cs.hku.hk/fyp/2016/fyp16019/

7 Challenges and Mitigation

The potential challenge of the project lies in the inaccuracy of the models chosen. For instance, the model developed might only be applicable for a small range of data. To minimize the error, we will use the iterative implementation method, which will focus on implementing one function at each iteration. Each iteration will include testings on different sets of data and revisions to make sure that the functions implemented meet the needs.

Another challenge is the insufficiency of the data available. At this moment, we only have access to the admission data of the past three years, which might create limitation in the model we build. To address the problem, we will make the system scalable enough to take in new admission data each year such that the models can be easily adjusted to be more robust.
8 Schedule

<table>
<thead>
<tr>
<th>Date</th>
<th>Task</th>
</tr>
</thead>
</table>
| September       | • A detailed project scheme  
                  | • Project Web page                                                   |
| October         | A prototype with the following implementations:                      |
|                 |   • one basic data visualization tool                                |
|                 |   • one basic data analysis tool with a specific data mining algorithm |
| November to January | • A prototype with all the determined data visualization tools  
                        and data analysis tools                                               |
|                 |   • Interim report                                                   |
| February to March | • A web application with the following functions implemented:     |
|                 |     – Smart Filtering                                               |
|                 |     – Intelligent Matching                                          |
|                 |     – Outcome Prediction                                            |
|                 |   • Integration of the web tool into the existing system            |
| April           | • Testing and evaluation of the web application                      |
|                 | • Final report                                                      |
|                 | • Project presentation                                              |
| May             | Project exhibition                                                  |

9 Work Division

<table>
<thead>
<tr>
<th>Task</th>
<th>Teammate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data visualization</td>
<td>Xu Fangyuan</td>
</tr>
<tr>
<td>Data analysis</td>
<td>Wu You</td>
</tr>
<tr>
<td>Prediction</td>
<td>Xu Fangyuan</td>
</tr>
</tbody>
</table>
10 Conclusion

This project aims at addressing the problem of inefficiency and inaccuracy in the existing admission process. To solve the problem, the objective is to develop a data mining web tool that facilitates the extraction and analysis of the admission data of HKUCS graduate programs. With the functionalities provided by the web tool, the faculty members will be able to make prediction and gain insights that are meaningful for the admission process and strategies. As the problem is common in recruitment process, we believe that the project is scalable and can be extended to other recruiting practice.

References


