International School of **Engineering**
Chulalongkorn University

DATA VISUALIZATION FOR CHULALONGKORN UNIVERSITY

Assanee Sukatham, 5531334621, assanee.su@gmail.com, 0847074536

Krerkkrai Thamjarat, 5531216221, tino.thamjarat@gmail.com, 0809707776

Jarindr Thitadilaka, 5531219121, jarindr23@gmail.com, 0863922333

Advisor: Asst. Prof. Athasit Surarerks Signature Date

PROPOSAL

TABLE OF CONTENT

1. Introduction

Chulalongkorn University has a lengthy history. Since founded in 1971, there are a lot of people from different parts of the country that have been part of Chula. There are approximately 10,000 students graduated each year which each individual has records of their profile and performance during their time in Chula kept in the Office of Registrar database. Not only students have their records but also instructors and staffs in Chula. Those raw data was kept in the registration department which could be use to further improve strategic planning for Chula. However, these data have not been put to use yet.

Our university have “raw data”, also known as primary data, which are collected directly from sources. Raw data has no sense and meaning. They have to be processed or manipulated in someway in order to get meaningful information. Those refined information will be a valuable asset of our university to use or to present them to the world community.

Currently, Chula has no refined data and no official channel to represent them to outside entities who have an interest in Chula. For example: The World University Ranking Organisation, prospective students and staffs, alumni and etc. Raw data from the office of registrar could play a vital role in the mining process to extract the information that those users need and further enhance Chula reputation and provide an up-to-date information in multiple aspect for both people of Chula and people outside.

Information is not our only point of concern, UI/UX also play a big part in this project. Having a valuable information means nothing if they’re too complicated for the user to take in, it will be just texts and numbers. Therefore the way information represent need to be simple and attractive in order to provide the best user experience for visitors.

For all the reasons we previously stated, we would like to propose a project to create a web application that provide information of Chula in every possible aspects and improve decision making for Chula registration department and provide information for interested people.

2. Background

2.1. Required Tools

2.1.1. Designing Tool

1. Sketch 3
2. Adobe Photoshop CS6
3. Adobe After Effect CC 2014
4. InVision

2.1.2. Software Developing Tool

1. Web Development Tools

Sublime Text 2

It is a cross-platform source code editor with a Python application programming interface (API). It natively supports many programming languages and markup languages, and its functionality can be extended by users with plugins, typically community-built and maintained under free-software licenses.We will use sublime text as a default source code editor due to the light-weightness of the program and the flexibility of installing plugins.

WebStorm

The smartest JavaScript IDE WebStorm is a powerful IDE, perfectly equipped for complex client-side development and server-side development with Node.js. WebStorm’s smart code editor provides first-class support for JavaScript, Node.js, HTML and CSS, as well as their modern successors. Take advantage of code completion, error detection, refactoring and more. Because sublime doesn’t enough capability for debugging and refactoring.

HyperText Markup Language (HTML)

Commonly referred to as HTML is the standard markup language used to create web pages. Since the project is to develop a web application, it is the domain knowledge that developer have to immensely concern.

Cascading Style Sheets (CSS)

CSS is a style sheet language used for describing the presentation of a document written in a markup language. Although most often used to set the visual style of web pages and user interfaces written in HTML.It is used to define styles for your web pages, including the design, layout and variations in display for different devices and screen sizes.

Bootstrap

Bootstrap is a front-end framework from Twitter designed to kickstart the front-end development of webapps and sites. Among other things, it includes base CSS and HTML for typography, icons, forms, buttons, tables, layout grids, and navigation, along with custom-built jquery-plugins and support for responsive layouts. Bootstrap will help in styling the webpage easier, reduce self writing of CSS files to handle responsive layout of our project.

Javascript

JavaScript is a programming language used to make web pages interactive.It A high level, dynamic, untyped, and interpreted programming language.It has been standardized in the ECMAScript language specification. Alongside HTML and CSS, it is one of the three essential technologies of World Wide Web content production. Javascript is a client side programming language which will handle the interactive part of our website.

Angular.js

It is a Javascript framework which It lets you use HTML as your template language and lets you extend HTML's syntax to express your application's components clearly and succinctly. Since our website will mainly be dynamic which Angular.js will play an important role, reducing the code that need to be written in javascript and create a more structured coding style.

Node.js

Node.js is an asynchronous event driven framework, Node.js is designed to build scalable network applications..It is an I/O environment built on top of Google Chrome's JavaScript runtime.Essentially, a server-side implementation of JavaScript.It’s event-driven I/O model makes it easy for developers with JavaScript knowledge to build high-performing, scalable, and highly concurrent web applications.Node.js will mainly be the project’s backend which will handle I/O of the system.

Express

Express is a minimal and flexible Node.js web application framework that provides a robust set of features for web and mobile applications.It enable more function for node.js, reduce time of coding,improve overall function for node.js which is our main backend system,

MongoDB

A cross-platform document-oriented database. Classified as a NoSQL database, MongoDB the traditional table-based relational database structure in favor of JSON-like documents with dynamic schemas.MongoDB will be the database system of the project, which make future implementation that involve scalable database system not a problem, MongoDB also works well with Node.js which is the main backend of the system.

2. Data Mining Tool

RapidMiner Studio

RapidMiner is a free software platform produced by RapidMiner Inc, intended to use for data analytics domain such as data mining, machine learning, text mining and predictive analytics. Large volume of data can be analyzed by using different data mining algorithm provided by RapidMiner (e.g. Statistical classification, regression). The output data from the mining process can be visualized and further utilized by other processes (such as Linear Programming). RapidMiner is an open-source software and thus it is free to use. In contrast to commercial data mining software (IBM SPSS), RapidMiner provides sufficient functionalities for mining a considerable size of data set with a multi-dimensional data attribute

2.2. Required Knowledge

2.2.1. Design Knowledge

1. UI Theory

Responsive

In today's world, the use of mobile devices including tablet increase dramatically. Unfortunately most website is not optimized for the devices and can create frustrating feeling for the user. Mobile devices are often constrained by display size and require a different approach to how content is laid out on screen.

There is a multitude of different screen sizes across phones, "phablets", tablets, desktops, game consoles, TVs, even wearables. Screen sizes will always be changing, so it's important that your site can adapt to any screen size, today or in the future.

Responsive web design, originally defined by *Ethan Marcotte* in A List Apart responds to the needs of the users and the devices they're using. The layout changes based on the size and capabilities of the device. For example, on a phone, users would see content shown in a single column view; a tablet might show the same content in two columns.

Squint Test

Squint test is a test that the the website will be blurred out to test the visibility when user scan the website.When blurred, is it still be able to distinguish the element that should stand out. Are there too many elements on the page which could confuse the user. The main point of the test is to get a high-level view of the visual hierarchy of your work.

2. Tone & Theme

Flat/Material Design

Flat and material design is the trend of today's design. Flat design was first introduced by Microsoft's Metro UI, which came into existence with the launch of Windows 8. And for material design was first introduced by google. Both of them are quite similar to each other.

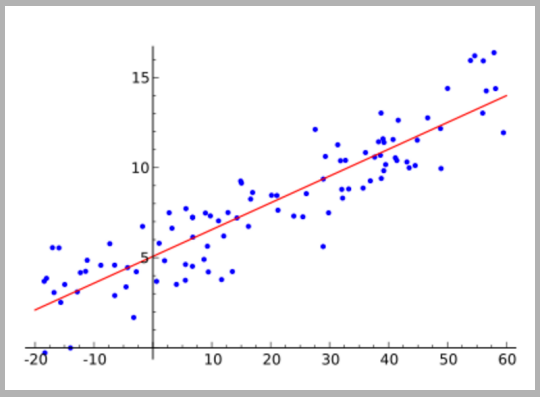
Basically, flat design is the complete opposite of skeuomorphic design, mainly due to the lack of all the stylistic or 3D objects. Material design, on the other hand, keeps some stylistic elements but not the skeuomorphic ones, yet it also retains the minimalistic appearance of a flat design. Material design is a flat design that play with some z-index to stylish the content to have some 3D aspect.

2.2.1. Machine Learning Knowledge

Regression Analysis

Figure 1

Regression Graph

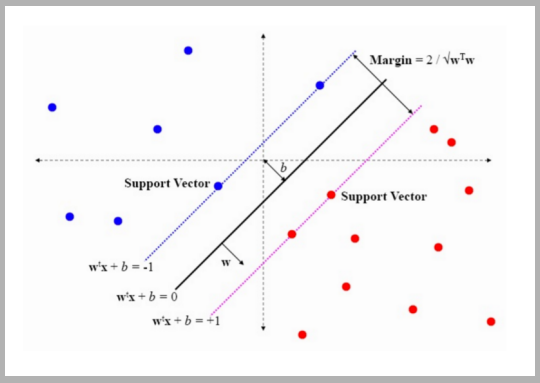


Regression analysis is a machine learning algorithm that is used to analyze the correlation of two or more variables. The algorithm is based on the statistical linear regression approach. In the Simple Linear Regression, data is modeled as a relationship between one predictor variable X (independent variable) to an output (dependent) variable Y. Linear Regression finds the most optimal straight line (in case of linear regression analysis) from the given set of independent and dependent variable (X,Y) with different prediction errors. The regression line is defined as a line taking the least sum of the square of the prediction errors.The line is referred to predict the value of the dependent variable Y by input an independent data (X). The precision of the output depends on the correlation among the variables. The regression algorithm maps the input data to the infinitely-possible continuous output Y.

Support Vector Machine

Figure 2

Support Vector Machine Graph



Support Vector Machine is a supervised machine learning algorithm that is used for linear classification and regression type data mining. The algorithm tries to identify unknown-type of data into classified group by using a training data set, finding adjusting the most suitable straight line (linear kernel) that separate groups with the largest space between the nearest member of the group to the median line. The algorithm learns how to classify an input X into an output Y by inputting the training data (x,y)S. SVM makes a decision according to the information or knowledge (training data) it currently holds.

Classification

Classification algorithm is one of the supervised machine learning algorithm. The algorithm works to separate and identify different unknown types of data into sets of different categories, inputting a predictor-independent variable X to correlate with a set of finite-discrete output variable Y.

2.2.2. Data Mining Knowledge

Association Analysis

Association analysis is the method used to find the underlying relationship among attributes in the large data set. By scanning through the large set of data, the algorithm determines the association rule by counting how likely the data set X induce the data set Y. The goal is to generate all possible association rules that suggest the relationship among the attributes.

Association rule

$$X \rightarrow Y$$

If data set X occurs, it has a high chance that the data set Y will occur.

For example:

$$\{\text{Architecture, Male}\} \rightarrow \{\text{Late payment}\}$$

This suggests that male architecture students usually pay the tuition fee late.

Support count

σ(x) = Number of transactions that contain x

Support

Support is the minimum threshold to exclude the association rule that might occur by chance, i.e. the frequency of the occurrence of the association rule transaction, with respect to the total number of transactions, does not reach minimum value of support count to be considered as significant.

Support(X→Y) = (XY)N , N is the total number of transaction

Confidence

Confidence is the probability that the association rule will hold by taking a ratio of the number of the transaction that contain X and Y over the number of transactions that contain X.

Confidence(X→Y) = (XY)(X)

Confidence measures the consistency of the association rule whether to what degree the data set X certainly induce the data set Y. The association rule that does not reach the minimum confidence threshold will be excluded from generated association rules. It is called minimum confidence level (minconf)

For example:

$$\text{Confidence}(\{\text{Architecture, Male}\} \rightarrow \{\text{Late payment}\}) = 0.6,$$

This convinces that there is 60% chance that male architecture students will pay the tuition fee late.

Association Rule Generation
The objective of the association analysis is to extract the strong confidence association rule from the itemset that surpass the minimum support constant. Association rule represents the strong reference from itemset X to itemset Y. Therefore, we can infer the hidden relationship between attributes by finding these association rules.

3. Literature Review

3.1. For Designing

Frightgeist Website

https://frightgeist.withgoogle.com/

It is a website that shows the statistic of Halloween costume worn in United States in the 2015. It is a great example for data visualization. It uses a lot of technique such as Geographical Graph, Bar Graph and many other technique. The design is very simple and clean. It is also responsive which still have the same functionality as the full desktop site.
We use this as an example for data visualization and responsive of our website.

Aprilzero Website

http://aprilzero.com/explorer/

Aprilzero is a website that shows real-time statistics of a man name Anand Sharma. It gets data from his mobile phone Gyroscope and display it on the website. The website displays various kind of data such as his workout statistics by FitBit app, his latest location by Foursquare, his latest photo by Instagram and many other interesting data. The design is gorgeous with simple yet interactive UI. You could hover around and everything on the page is interactable.
Our design will be inspired by this website and we will provide similar interactiveness for the user.

3.2. For UI Theory

Keystrokes Level Model

KLM predicts task execution time from a specified design and specific task scenario. Basically, you list the sequence of keystroke-level actions the user must perform to accomplish a task, and then add up the times required by the actions. It is not necessary to have an implemented or mocked-up design; the KLM requires only that the user interface be specified in enough detail to dictate the sequence of actions required to perform the tasks of interest.

User-friendly has been a term of non-functional requirement that is hard to measure numerically. KLM propose a way for developers to estimate the execution time of tasks in their UI. We will use this theory as a criteria to measure the performance of our UI.

4. Objective

We aim to build an informative website of Chulalongkorn University for both outsider and Chula people by mining interesting information regarding Chula students from registrar office's raw data and visualized it on the website.

5. Methodology

5.1. Software Development Life Cycle Approach: AGILE

Agile methodology is chosen because Chulalongkorn University does not have data visualization website before therefore we have zero knowledge of user requirements and we also have time limitation hence we cannot do market research. We want to focus on building the website and get feedback from users as fast as possible so we could deliver the right product for them. Moreover, due of the nature of senior project. There is a bi-weekly progress report and meeting. An iterative nature of AGILE is a perfect characteristic to bring significant progress to present to our advisor.

We will use “BamBam!” as our project management tool. BamBam! is an online Agile project management tool which is free for small team under 10 members. We will use it to streamline the workflow and also to communicate among our team. Our project link is at:

- https://chulayolo.dobambam.com

We use GitHub for website version control and slack as our document version control.

- https://cudatasite.slack.com
- https://github.com/vtno/senior_project

We will use trello as our main SCRUM tool.

- https://trello.com/b/O7a94XuR/cu-data-visualization

5.2. Technical Approach

5.2.1. Programming Languages:

Our website is separated into two parts which are front-end and backend. For front-end development we will use the following languages:

1. HTML5
2. CSS3
3. JavaScript

HTML5 and CSS3 is the latest version of web structure language which is design to support more dynamic and responsive website trend that our website will follow. JavaScript is chosen as our scripting language for client side because our team have been exposed to JavaScript and are familiar with the language.

For backend development we will use the following languages:

1. Node.js
2. MongoDB

We decided to use **MongoDB** as our database because we aim to have a scalable system that could handle dynamic transaction in the future which MySQL is not suitable for that kind of system. **Node.js** follows through because it is more compatible with MongoDB than PHP. Also, Node.js handles socket programming better than PHP from our experience with both languages.

5.2.2. Programming Language Frameworks

1. Bootstrap

We choose Bootstrap as our front-end framework because it is the most popular framework for responsive website which is easy to use. We also have lots of experience with Bootstrap.

2. Angular.js

Angular.js is chosen because it is a very powerful tool to handle html syntax and create stunning visual animation for our website.

3. Express

We use it as a framework for Node.js which provide more functionality for Node.js.

5.2.3. Developing Tools

1. **Sublime Text 2** - Main text editing tool
2. **WebStorm** - Text editing and web debugging Tool
3. **Microsoft Excel** - To edit, view, process and convert raw data.
4. **GitHub** - For version control

5.2.4. Data Mining Tool

RapidMiner

RapidMiner is interesting for our project because it supports many different data mining models that are suitable for analyzing Chula data which have different type of data attributes. Classification, regression, clustering, segmentation and association algorithm are needed according to the type of problem. RapidMiner is famous in sense that it offers a powerful and customizable visualization options for both training set and outcome. Charts and statistical views are applicable to the proposed data visualization website. It also comes with numerous useful features such as third-party data source support (e.g., Excel, IBM SPSS, and MySQL), custom R and command line script. Mining algorithm for Chula data can be fully customized for suitability by writing extra scripts. Besides, RapidMiner Studio is free to use on the starter and personal level.

5.2.5. Designing Tools

We chose Sketch as our designing tool because our designer has a lot of experience with it and also it is compatible with our prototyping tool. It is easier to setup a workflow than using other vector-based designing program like Illustrator.

Adobe Photoshop CS6

It is the best image editing software in the market. We chose this as our image customization tool to make icon and manipulate photos.

InVision

InVision is a free web-based prototyping tool with collaboration function which is perfect for prototyping our website because it is very easy to comment on the project and make correction on the fly. Also, it provide sync system that automatically sync Sketch file into the prototype.

5.2.6. Developing Environment

Hardware Requirement

- Macbook Pro 15' Retina Late 2014
- ASUS RoG G550JK
- Dell Venue 11 Pro 7140

Browser Requirement

We will develop our website for the latest version of 4 popular browsers which are

- Google Chrome
- Mozilla Firefox
- Safari
- Microsoft Edge/Internet Explorer

*For mobile site, we will use emulator on Google Chrome

Operating System Requirement

We will develop our website on OSX El Capitan and Windows 10

5.2.7. Functional Requirements

1. User System
2. Database System
3. Data Visualisation

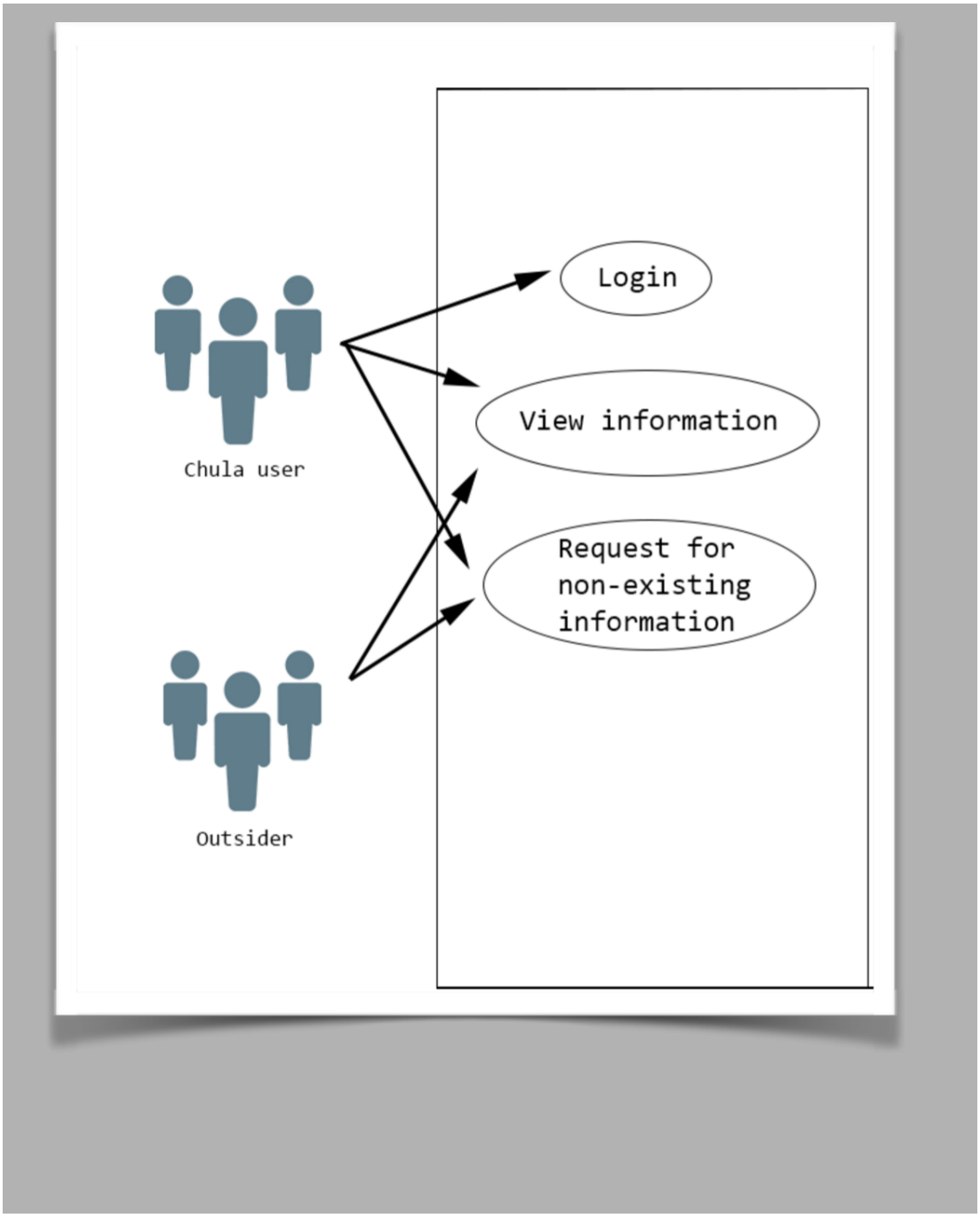


Figure 3

Use Case Diagram.

5.2.8. Non-Functional Requirements

1. User Friendly
2. Fast query
3. Responsive

5.2.9. Acceptance Criteria

Functional Criteria

1. Successfully log in/out of the system by Reg Chula account
2. Visualise correct data from the database
3. Each page is displayed correctly according to the sitemap

Non-Functional Criteria

1. Displaying the right data for user in less than the passing criteria for each task. The result will be measured by KLM.
2. Each query from the database must not take more than 2 second to display
3. The website is displayed correctly on regular screen, mobile and tablet

5.2.10.Constraints

Data Constraints

Reg Chula data dated back to only in 1995 so we cannot provide statistics before 1995. Also some data is incomplete, it is not possible to forecast/predict every interested relation.

Hardware Constraints

Could we use Chulalongkorn server as our server for the project when it is officially launch.

Time Constraints

Due to time constraints, we cannot do deep market research to find our user requirement for the website. We will use agile to help us see the clearer requirement in each sprint phase.

5.2.11.Stakeholders

Outside Users

We classify outside users into 2 types which are:

1. Prospective Students

This segment will need data such as admission score statistics, high-school statistics, jobs after graduation statistics and other information regarding Chulalongkorn University or individual faculty. The data provided must be in the right format for high school student to comprehend.

2. Other interested people

This segment will concern only statistics of Chulalongkorn University as a whole such as which province does most students came from, abroad studying students data, scholarship and internship statistics and etc. This segment also include organisation that needs Chula data for annual ranking of world or national university ranking. The data provided for this segment must be informative and on point with their interest so it could improve Chula ranking or attract outsider interest.

Chula Users

This segment concern internal usage of the website by Chula people. The data displayed will be more niche such as which Gen-Ed course has the best chance to get an A, Is high school grade related to college performance, which faculty has the highest GPA and many more. We need to use more raw data to do the mining than the dataset for the outside users.

6. Project Detail

6.1.Required Data

This project will use the given set of data , primarily from the Office of Registrar and any additional information provided by other divisions, with the data mining tools to observe patterns and relationships between attributes. The followings are the required data for the mining process, mainly the information about Chulalongkorn's students, personnel and classes.

6.1.1.Required Fields

Personal Record

- First name, Last name
- Student ID
- Academic record (by semester)
- Grade (by subject)
- GPA
- Address
- Secondary school
- Faculty and program

Course Information

- Course title
- Class date and time
- Instructor's name
- Number of registered students
- Number of total seats
- Withdrawal record

Annual admission data

- Average (High school) GPA
- School name
- Average admission score (by program)
- Number of acquired students

6.2. Relations and Visualisations

These are example of relations we are interested in:

- Number of students in each year (by faculty/program)
- Number of students by sex
- Number of students by secondary school
- Subjects that are attractive for students
- Withdrawal according to subjects and instructors
- High school relation with faculty
- Admission score trends
- Students distribution by district/province
- Admission score in relation with faculty selection
- Grade according to instructors
- Relations among secondary school GPA, faculty and current GPA
- Average grade and withdraw behaviour

Benefits

1. Current students
   - Subjects that are attractive to students
   - Withdrawal according to subjects and instructors
   - Grade according to subjects and instructors
   - Average grade and withdraw behaviour

   Student can access information about classes. Normally, these information are not available to them. They can use the information to make a study plan.

2. Prospective students
   - Admission score trend
   - Admission score in relation to program selection
   - Relations among high school GPA, faculty and current GPAX

   Prospective students can access information about admission score history and trends. They can use the information to make decision about faculty and program application.

3. Chula
   - Number of students in each year (by faculty/program)
   - Number of students by sex
   - Number of students by high school
   - Student distribution by district/province

Chula organisation and personnel can use these information to help in the strategic planning.

7. Scope of the Project

7.1. Deliverables

We need to provide two deliverables in this project

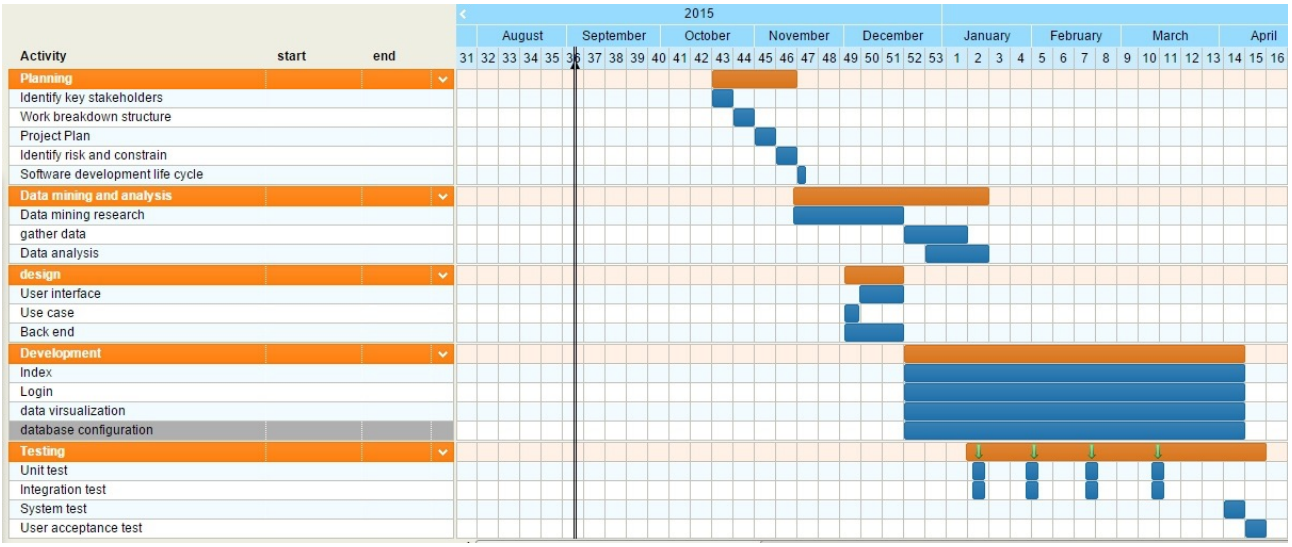
1. Fully functional website
2. Processed information from data mining

7.2. Limitation

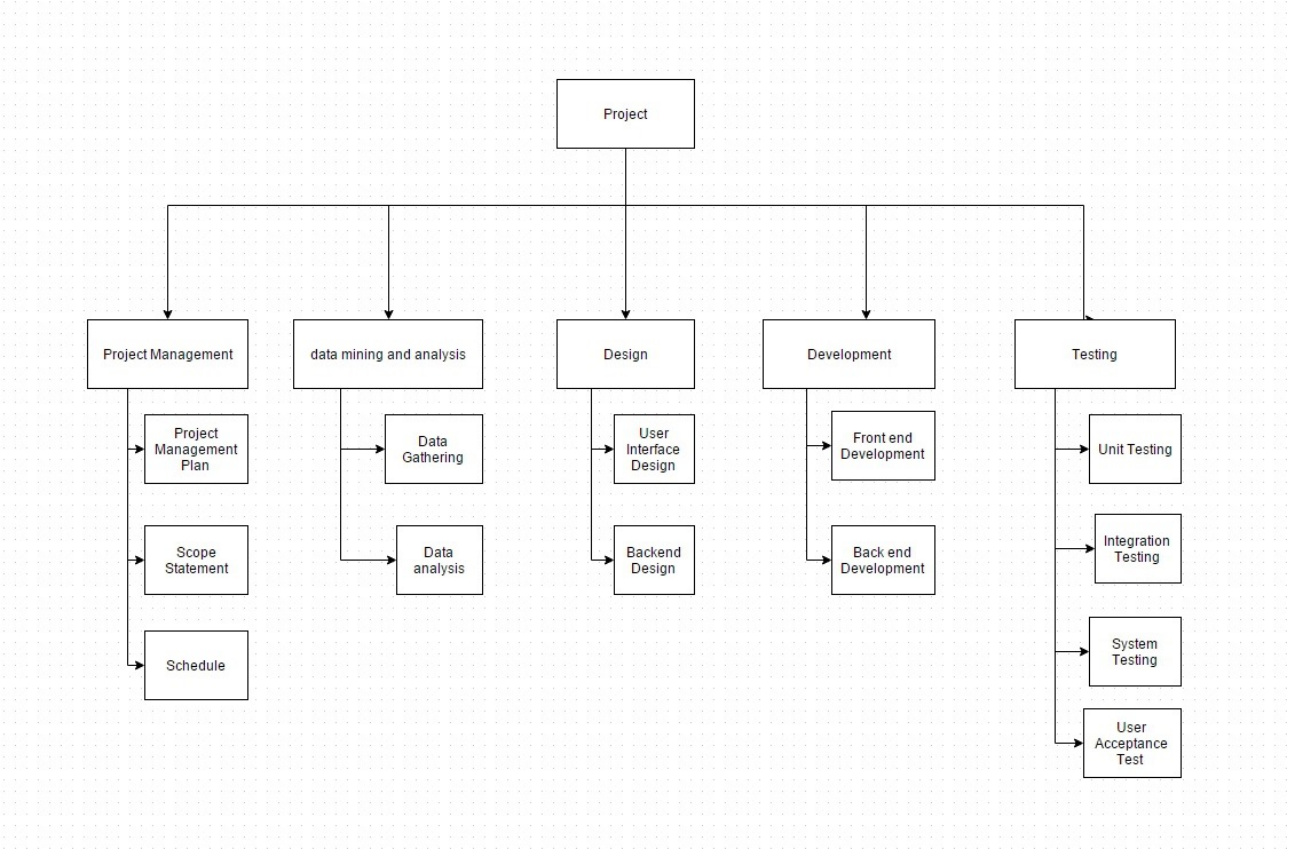
1. The website will not be real-time mining service. We will separately process raw data and upload them into our database hence users cannot query for a specific information that we have not process. We will use email list system to find new question that users want to know, try to find the answer and upload them onto our website.
2. Our website will not provide sensitive data such as individual address, name or age.
3. We only consider numerical data for mining not including text (text mining)

8. Plan and Schedule

8.1. Gantt Chart



8.2. Work Breakdown Structure



8.3. Team Members

Jarindr Thitadilaka - Jaja

Our full-stack developer who will be responsible for backend and frontend development of the website.

Krerkkrai Thamjarat - Tino

Our project manager and UI/UX designer who will be responsible for UI/IX design, data visualization and keeping agile project management on track including activity like SCRUM and Sprint.

Assanee Sukumtham - Hades

Our data analyst who will be responsible for data mining and machine learning to process raw data and get the desired information to display on our website.

9. Expected Outcome

The expected outcome of the project is a fully functional website which could provide information for Chula User and Outsiders. The website will be viewed by people that have interest in Chulalongkorn University. Providing data that people usually want to know but not exist in the official Chulalongkorn website.

10. Benefit of the Project

The website will be benefit mainly to Chulalongkorn University. One main purpose is for easier to find information to rank the Universities.Chula users could easily find the frequent ask questions about the subjects and grade in their faculty and the outsiders could look for the information about Chulalongkorn University's Faculties and Chulalongkorn itself.