



UPPSALA
UNIVERSITET

IT 11 085

Examensarbete 30 hp
November 2011

Process-Oriented User Behavior Study Based on Machine Learning

Yuting Wu

Institutionen för informationsteknologi
Department of Information Technology



UPPSALA
UNIVERSITET

**Teknisk- naturvetenskaplig fakultet
UTH-enheten**

Besöksadress:
Ångströmlaboratoriet
Lägerhyddsvägen 1
Hus 4, Plan 0

Postadress:
Box 536
751 21 Uppsala

Telefon:
018 – 471 30 03

Telefax:
018 – 471 30 00

Hemsida:
<http://www.teknat.uu.se/student>

Abstract

Process-Oriented User Behavior Study Based on Machine Learning

Yuting Wu

With rapid development of network technology and an increasing number of users, web applications become more interactive and collaborative. Under Web2.0, it is much more difficult for users to locate required information from a large amount of data. This thesis concentrates on techniques and methods that analyze users' behavior and capture their requirements on the Internet, which in turn provides users with an efficient way to retrieve information from the Internet, and also a better personalized service according to their preferences.

After acquiring status quo of web applications, the thesis proposes an approach for user behavior analysis based on machine learning and process modeling.

The thesis studied the behaviors of students who want to apply for a foreign college and study abroad. On the basis of Vertical and Horizontal Model, the period of application is divided into four phases: Language Examination, College Application, Visa Application and Private Affairs Preparation. Relevant activities of each phase are modeled as Process with WebML. Data Model was constructed for clarifying the relation between data objects and Navigation Model was adopted for guiding navigation design for specific user.

In this thesis, algorithms in machine learning are investigated. Through taking advantage of classic Decision Tree and Naïve Bayes, data collected from survey is analyzed and modeled in order to summarize rules between them. Furthermore, an improved prediction model based on Confidence is built to speculate user's behavior. At last, validity of the approach was conducted by using a prototype system. At the end of the thesis, further works about the problem are discussed.

Handledare: Qin Liu
Ämnesgranskare: Ivan Christoff
Examinator: Anders Jansson
IT 11 085
Tryckt av: Reprocentralen ITC

Content

1.	Introduction	5
1.1	Background	5
1.2	Objectives.....	5
1.3	Fundamentals.....	6
1.3.1	User behavior	6
1.3.2	Process Analysis.....	7
1.3.3	Machine Learning.....	7
1.4	Study Content.....	7
1.5	Thesis Structure.....	8
2.	Behavior Analysis Approach based on Process and Machine Learning	9
2.1	Process Model.....	9
2.1.1	Concept Extraction	10
2.1.2	Vertical and Horizontal Process Model.....	11
2.1.3	BPMN	12
2.2	Data Properties Analysis.....	13
2.3	Prediction Model.....	14
2.3.1	Classification Algorithm.....	15
2.3.2	Algorithm dependent on Confidence.....	17
2.3.3	Standard for Comparing Algorithm	18
2.4	Hypertext Modeling	19
2.5	Summary	19
3.	User Behavior Analysis based on Student Platform	20
3.1	Process Model.....	20
3.1.1	Concept Analysis	20
3.1.2	Horizontal process modeling.....	21
3.1.3	Vertical Process Model.....	23
3.1.4	Collaborative Process Model.....	25
3.2	Data Model.....	26
3.3	Prediction Model Based on Machine Learning	28
3.3.1	Data Resource	29
3.3.2	Analysis tool	30
3.3.3	Data Collection and Analysis	32
3.3.4	Prediction Model based on C4.5 and Naïve Bayes	34
3.3.5	Improved Prediction Model based on Comparison of confidence.....	37
3.4	Hypertext Model	38
3.5	Summary	40
4.	Prototype System based on User Behavior Analysis	42
4.1	Architecture.....	42
4.1.1	System Module.....	42

4.1.2	Network topology.....	43
4.2	System hardware and software environment.....	44
4.3	User Behavior Capture Module.....	44
4.4	Experiment.....	46
4.4.1	Original User Data	46
4.4.2	Experiment Result and Analysis	47
4.5	Summary	48
5.	Conclusion.....	49
5.1	Results	49
5.2	Future work.....	49
	Acknowledgements.....	50
	References.....	51
	Applendix A Distribution of Naïve Bayes.....	53
	Applendix B Implementation of Improved Algorithm.....	56

1. Introduction

1.1 Background

With the advent of Web 2.0, web-based application is seeing great changes with more interactive information sharing and interoperability. User-centered design is applied. The 25th Chinese Internet Status Report showed that the number of Internet users had reached 0.384 billion and the growth rate was raised to 28.9% until December 2009. The excessive amount of data in the Internet makes the information retrieval a hard problem to solve. For users, it is hard to find out the information they need from the Internet. How to help users efficiently locate the information has become an interesting research topic.

Most of the Chinese websites arrange information in a mechanical way. The articles are listed to their readers according to their popularity, which cannot satisfy most of the users according to a recent research. The interest and requirement vary from user to user, since each user has different background and interests. Moreover, some information supply websites such as Baidu is based on Bidding Rank, which violates the user-centered design guideline. Recently, an increasing number of user behavior studies are conducted, because users' potential requirement can be depicted by learning their behavior.

Machine learning is the scientific discipline concerned with the development and design of algorithms that identify relations between data. Human behavior is recorded as a series of historical data with attributes. Then derive general principle of action which is called as process. However, human behavior is so complicated that either process or algorithm fails to capture its model alone. Thus, improved algorithm based on process is necessary for special situations

1.2 Objectives

After a long period of accumulation, human beings could produce a series of sequential activities for certain tasks when they were studying, working and living, which is called Process. It turns out that human beings make decisions in certain activities getting accustomed to certain conditions or circumstances. Therefore, defining human being's behavior as processes will contribute to user behavior analysis and prediction.

This thesis studies the group of students who want to study abroad and proposes a method of refining user behavior based on process analysis and machine learning. The concept of vertical and horizontal flow is applied to define users' phases in the process of applying a foreign college and detailed activities on each phase. It sums up users' different orientations for the same class of objects during each activity through using the data collected from the behaviors of the students. In order to fulfill above targets, C4.5 and Naïve Bayes algorithm are applied to predict users' behavior. The prediction model is improved by Confidence, which is the

measurement of whether the prediction can be convinced and processed. This approach is designed to make the information platform serve students in the light of their preference, and avoid the interference from the excessive amount of data. Meanwhile, a prototype system, which is applied in the student studying aboard platform, is implemented to verify the feasibility and suitability of this approach.

1.3 Fundamentals

1.3.1 User behavior

It is well-known that each individual has his/her own personality. More specifically, personality is personal features established and developed under certain social environment and educational background. Therefore, everyone differs in terms of psychology, behavior, physiology, personality, strengths, interests and values. This explains the reason why people will have their own distinct requirements. Generally speaking, it is hard to conclude single and common rule of behavior superficially within short term. But some patterns for certain kind of persons can be obtained after a long term observation.

Traditionally, approaches of capturing user behavior including survey, interview, group discussion, experiment, observation, record and so on are adopted commonly. However, nowadays data mining has been introduced and converged with them. Then new approaches can be mainly classified into following two groups:

1. User Characteristics Analysis

User Characteristics Analysis is to discovery specialties on behavior of various users. It is a prerequisite of providing users suitable services, because of its importance to obtaining users' preference. For instance, people residing in the different district will choose various tutorial classes according to their location, e.g. people in Shanghai will not choose classes carried out in Guangdong because of the difficulty arising from the distance. After location analysis, the system generally filters some candidates have distance problems and provides more feasible services for users, during which users' properties are screening to understand the relation between them, in order to speculate next behavior.

2. Classification and prediction

Users can be classified into different categories by some classifications [1]. For example, by analyzing weather, humidity and temperature before trip, it speculates whether the person could make an out-going. If acceptable, it provides the traveler relevant services in advance. In this thesis, activities of the process are set up as objects of classification after taking users' properties into account. Then, various classifications are conducted to deduce the possible next activities for the specific user.

1.3.2 Process Analysis

Process Model [2] concludes users' normal behavior patterns, which is the foundation of investigating properties and provides input parameters for constructing prediction model in Machine Learning. Process Model is divided into three phases: Concept Meta Modeling, Process Modeling and Implement Modeling, which describe the process from the perspective of what will be involved, how to build and how to implement respectively. Depending on Process Model, users' potential requirements are captured by probing historical data.

Business Process Management [3][4](BPM) plays an important role in developing Business Application. At present, via the BPM modeling, business process modeling notation [7](BPMN) is a popular symbol adopted by today's process design tools. Moreover, as an excellent symbol for Business Process, BPMN is able to depict the human behavior process properly as well.

1.3.3 Machine Learning

Machine Learning (ML)[8] is the scientific discipline that studies how computer simulates or implements human's behavior, in order to acquire new knowledge or skills and improve its performance by reorganizing the existing knowledge structure (model). In this aspect, this thesis adopts two algorithms: Decision Tree (DT) and Naïve Bayes (NB) algorithm.

Classification, Association Rules, Clustering are typical examples of ML. As for Classification, DT is one of them, which builds a logic tree depending on training set. ID3 [9][10] and C4.5[11] are improved DTs. Moreover, DT is applied in so many expert systems like MORGAN-a gene identification system, Network Security [12] and Peer-to-Peer system [13]. For example, C4.5 is even used to predict human talent [1]. The Association Rules [14][15] investigates relation between properties of data set to generalize frequently appeared rules which are valuable. Main algorithms have Apriori and the improved FP-Growth algorithm. Customer' buying behavior is a key field in which Association Rules is involved.

1.4 Study Content

This thesis was intended to efficiently locate the information for users who are stuck in a flood of information, and proposed a process-oriented user behavior analysis method based on Machine Learning.

By studying the ordinary process model, it turns out that process model gives a general pattern of behavior. But it fails to solve the situation in which there exist multiply choices in one activity in the meanwhile. It is rather difficulty for process to describe rules for that. Thus this thesis improved it by Machine Learning to obtain the most possible behaviors in the next step.

Data Model and Navigation Model are conducted on related data in the process of students applying colleges. The Decision Tree and Naïve Bayes were validated under the TP Rate, FP Rate, Precision, Recall, and F-Measure. The improved method was validated under 258 data from the survey.

Finally, a prototype system was implemented and further validated under the 15 users.

1.5 Thesis Structure

There are five chapters in the thesis, and the main content of each chapter is as followed:

Chapter 1 introduced the background and goals of the study; illustrated the significance of user behavior, process model and machine learning; listed the study content and thesis structure.

Chapter 2 elaborated the process-oriented user analysis approach based on Machine Learning. By the vertical and horizontal model, this chapter analyzed basic four phases and detail activities of them in order to form behavior patterns. First of all, data objects in each activity are extracted to conclude their properties and their inner links. Then, Prediction model was built on the historical data to speculate the direction of behavior. Finally, Navigation design and module division provide the foundation of prototype system.

Chapter 3 applied the whole approach to target group i.e. students studying aboard. At first, a survey was conducted to collect original data. Then, analysis tool –Weka’s functions were introduced. By comparing the Decision Tree with Naïve Bayes under TP Rate, FP Rate, Precision, Recall, F-Measure, it summed up the advantage and disadvantage of them under 258 samples from result of the survey. What’s more, this chapter processed the 258 samples by improved algorithm. The correctness of the improved algorithm was listed with former two.

Chapter 4 implemented the prototype system. In this system, the prediction module and behavior capturing module were illustrated, and the improved method was validated under the system.

Chapter 5 concluded this thesis and demonstrated some possible future orientations of this problem.

2. Behavior Analysis Approach based on Process and Machine Learning

The goal of this thesis is to predict users' potential behaviors. Thus, users' historical behaviors are collected and studied first. Certain people are investigated in order to provide better information to the user of the student platform. Certain data set are collected when they are performing some tasks or during some activities. All the future potential activities will be the foundation of providing services.

This chapter proposed a behavior analysis method based on process and machine learning. The four phases in this method are shown in Figure 2.1. According to the vertical and horizontal process model, basic stages in certain user behavior and different activities in the each stage are figured out, which constitutes the foundation of user behavior model. Then relevant data objects are deliberated, especially properties and relation between data objects in each activity of process, to produce some attributes which can be used as input parameters of Machine Learning algorithms. In the next step, historical data of user behavior is processed by algorithms of Machine Learning, in order to construct behavior prediction model and conjecture user's possible direction of next behavior. At last, Hypertext Model presents the abstract architecture of web application by analyzing the whole behavior period.

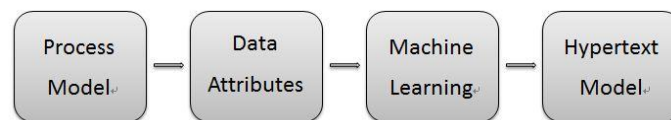


Figure 2.1 Analysis Process

2.1 Process Model

Process comes from what human being does in certain transaction and is the summary of classic experience. Each activity in the process represents a function module for a system. Traditional business software is produced based on the BPM [18]. Meanwhile, today's web applications are made in the definition of Business Process (BP) as well. However their BPs are more complicated. For example, in a group-purchasing platform, a customer will face thousands of candidates when selecting his/her favorite items. For each candidate, the customer has to do different operation. In China, two famous group-purchasing platforms-Meituan.com and Dianping.com, they are different in ordering process, such as the number of items they offer each day and the way of paying. Therefore, Process Model [16][17] is what has to be done in the stage of requirement analysis. Because it helps the architects get clearer idea of the system function. In this thesis, Process Model [18] is the first

step of analyzing user behavior.

2.1.1 Concept Extraction

Concept extraction is the first step of process modeling and works for clarifying the relevant major elements which guide the system design. In the process of Concept Extraction, what have to be declared are: Who are the potential users? What kinds of information and services users require? When users need those information and services?

1. User Object

User Object is target group in the whole analysis process. In either traditional software or web-based application development, it is important to set a user object correctly. User object guides the feasible project goal and determines methods of analyzing the whole process. For an online community application, workers can be defined as the target users (such as kaixin.com). Then trust relation between two users is influenced by the working relation between them. If the target is student (renren.com), relatively speaking, school relation (such as classmate, student-teacher) becomes measures. Those two situations will launch different services and community software.

2. Content

Services, which are defined as the functions of project, are the core of the entire project. When doing process analysis, analyst should clarify what services should be defined and what behavior should be modeled. She/he should conduct a research for serving objects (users) ,and then reasonably consider market conditions and objects' features to infer precompetitive functions in line with users' requirement. Extraction of service content is a process going from complex to simple. Distinct functions for each user are developed in different aspects.

In the beginning of project, superabundant function will harm the health of development because of the limited time and labor. The search engine like Google succeeds in helping user obtain right information in cyberspace at start. The new cloud service like Google Doc and Google Code were launched later. In the initial stage of a project, a number of classic and competitive functions should be introduced into market.

3. Phase Classification

Phase Classification is a significant approach for sorting out services. In a project, functions should be implemented are diverse. In order to better develop those functions and make them more conveniently for users to use, they should be classified at first. Just like some e-commerce websites, classify goods for buyer, in order to quickly locate helpful resources. NDT (Navigation-Driven Technique) conducts Navigation-Driven requirement analysis, proposes an approach with a start point of navigation and improves users' experience. Classification is one part of designing Navigation. This is a hierarchical method for process analysis. Services are

collated by phases, which in turn promote clear thought on process analysis.

2.1.2 Vertical and Horizontal Process Model

In this thesis, the process is classified as horizontal and vertical ones. Horizontal process is long-term and composed by users' different phases. Vertical process is the specific behavior activities in each horizontal ones.

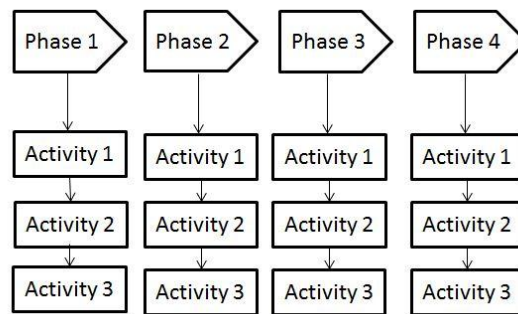


Figure 2.2 Vertical and Horizontal Process Model

1. Horizontal Process

Horizontal Process is made by different phases in users' behaviors. For example, student's studying process can be divided into phases of elementary school, middle school, high school, college and so on. In different phases, activities to be undertaken and things to be learned by students are different. Therefore, their requirements are various. The phase a user might be in must be clarified before further investigating that phase.

Each phase is implemented as a certain sequence. According to the procedure of human understanding things, each user in advance considers what will happen later. Thus, current and subsequent phase will be interested phases for that user. The relation between those two phases is study target for behavior analysis.

2. Vertical Process

Vertical process is a collection of the specific behavior activities of each horizontal phase and one kind of the latest behavior pattern. For example, a student who wants to select courses in college should experience the following process: logging in a specific website, finding out certain credit demand, browsing available courses, selecting courses and so on.

By summing up the series of specific activities, what can be obtained are:

- 1) Users' requirements on services, such as finding out necessary credits
- 2) Data objects should be cared, such as user, courses, credits and so on.

The main purposes of analysis of vertical and horizontal process are to extract the users' basic behavior patterns and related data objects and to provide a basis for data analysis later. It reflects general rules of human behavior. It is necessary to further deliberate and refine the direction of user behavior by machine learning.

2.1.3 BPMN

In this thesis, after concept extraction on process and specific activities analysis, the process model is depicted by BPMN. Process is composed of multiple elements, which can be described from different dimensions and perspectives, usually including functionality, business logic, organization, knowledge, goals, data and products. A business process model refers to network made up of graphic objects. Graphic objects include activities and flow controls which define the sequence of execution for these activities.








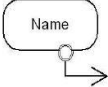





Events	 Start	 End	 Intermediate
Gateways	 Or gateway	 Xor gateway	 And gateway
Activity	 Name Activity	 Name Activity with event catching	Looping, ad-hoc, and compensation activities are provided
Flow	 Sequence flow	 Message flow	 Data Association
Grouping	 Pool	 Lane	Other grouping objects are provided: transaction, group, subprocess, ...

Figure 2.3 BPMN main elements

BPMN is a set of standards for defining business processes developed by BPMI (The Business Process Management Initiative) –www.BPMI.org. In Business Process Management, BPMN is used to define Business Process (BP). Nowadays, BPMN is shifted from 1.0 to 2.0. Compared to UML (Unified Model Language), BPMN provides a more specific and instructional business process analysis method and higher reliability for Process Model. Moreover, BPMN's graphic model for BP is simple and easy to operate.

The main concepts in BPMN:

Activity: The basic work unit of process. Usually it is operated by one user.

Constraint: Logic dependencies between activities, including Sequence, AND-split/AND-join, OR-split/OR-join, XOR-split/XOR-join.

Content of service is selected by Concept Extraction. Each activity is refined phase by phase. Then the direct-link relation between those activities is represented by Sequence. When there are multiply activities for selecting, alternative flow is used for depicting the relation between those activities by AND-split/AND-join, OR-split/OR-join, XOR-split/XOR-join. Because proposed user behavior analysis method is related to history data, message flow is applied.

Process Definition is able to clearly display the pattern of user behavior and

provides the foundation for later data properties analysis. Prediction model is based on machine learning. The relation model between data objects is conducted by statistics, information calculation and so on. According to this model, the corresponding prediction of future users' reaction in specific activities are made step by step. More detail historical data about the user behavior is so helpful for summing up the regulation and making the model. The effectiveness of prediction will be further improved by data accumulation.

2.2 Data Properties Analysis

Behavior analysis for users is aimed at exploring the most likely direction of users' behavior and helpful for users to retrieve required data quickly when they are browsing website. By analysis on process, what we obtained is specific groups of activities when users are performing a certain task or requirement. Meanwhile, data objects are produced from those activities. We can deduce the objects involved in the behavior process by process definition. Lots of judgment according to common regulation and experience are made on those objects properties, to describe part of the relation between them. So some Requirement Analysis methodologies are studied to speculate this relation.

Model-Driven [19][20] web engineering[21] method is highly abstract. It reflects the link and shift between data, showing the connection between the modules in the project. Most representative ones are WebML[22](Web Modeling Language) and UWE[23](UML Based Web Engineering).

UWE is a Model-Driven visual modeling approach, which is designed to improve the UML by adapting them to Web development environment. The model under UWE is considered and constructed from the navigation. So the idea is different from data-driven process model.

In contrast, WebML is designed as idea of considering the whole structure of web application from the data source level, in order to assist users better understand requirement.

WebML[24] consists four analysis prototypes: CommonElement, DataView, HypertextView and PresentationView.

CommonElement contains some core concepts for constructing WebML, such as DataType and Common feature. DataView defines some concepts about data like Entity, Attribute Relationship and so on. HypertextView defines a number of hypertext models. It includes some structurally related Pages, Content View and offers the combination relationship between them. PresentationView laies emphasis on views which will be finally displayed in the screen. Thus, WebML abstracts the hierarchies of the entire project framework and completes the final implementation by means of transformation

WebML, which pays attention to Data Centered website, contains Data Model, Hypertext Model and so on. It owns a set of completely steps from data to final implementation of sites. The analysis pays attention on the relation between data and derives modules of websites and their implementation.

Data Model [25][26] is one kind of typical Entity-Relationship Model, which can

transform the relationship between Entities into Concept Model. Entity is a data object, which distinguishes the objective existence of things, like User and Exam. Attribute plays a role of describing features of Entity, such as the name and age of User. Owing to system requirement, each Attribute has its own data type and feature.

In order to further discuss the prediction by machine learning algorithm, Data Model produces Attributes.

2.3 Prediction Model

It finds that, related activities of each activity for certain user may not be fixed. According to users' various properties, all subsequence activities change with different current situation. Those changes, by statistical calculation, can be obtained from a summary of some links between those activities. The above procedure forms prediction model.

This thesis aims at predicting most possible activities or content which user will choose. Machine learning has a large number of algorithms for mining the relation between data objects. So the model for the approach proposed in this thesis should be described as Figure2.4. The model adds a prediction module for suggesting rules between two data objects in order to classify the users' next behavior.

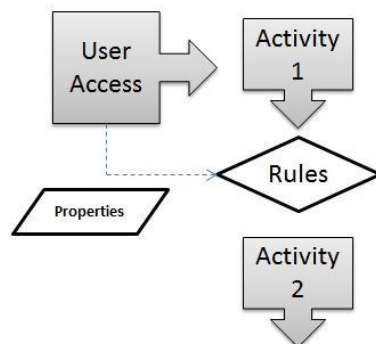


Figure 2.4 Analysis model

In the rule module (Rules in Figure 2.4), data objects related to the next activity can be obtained by analyzing the data attribute. Suppose we have a set of objects attributes $A = \{a_1, a_2, a_3, \dots, a_n\}$, a set of classification result $C = \{c_1, c_2, c_3, \dots, c_n\}$ and object instances

$$s \xrightarrow{\text{Rules}} c_i \quad (C_i \in C)$$

with attributes in set A, then once new object instance s occurs, the Rules module determines s which classification c_i belongs to. The rule module applies our improved algorithm for predicting, which explains in the next section 2.4.1 and 2.4.4.

Among classifiers, the two most widely used are Decision Tree Model and Naive Bayesian Model (NBC).

2.3.1 Classification Algorithm

C4.5 (Decision Tree) and Naïve Bayes are compared to identify which produces better prediction between activities, in order to pick up an efficient method of analyzing user behavior[27]. In problems of classification, usually an item should be put into one category. An item owns lots of attributes, which are regarded as vectors. i.e. $X = \{x_1, x_2, x_3, \dots, x_n\}$, X represents this item. There are numbers of classes, represented as $Y = \{y_1, y_2, \dots, y_m\}$. If X belongs to y_1 , X is marked by y_1 . This is called as classification.

1. C4.5

C4.5 [28] is put forward by Quilan for some problems occurred in ID3's practical application. So C4.5 is an improved algorithm for ID3. Both of them are heuristic algorithms based on information entropy.

1) Idea

In C4.5, some attributes of user behavior, which are obtained from Process Model and data property analysis, are set as tree node. Services required by users are regarded as tree leaf nodes. Different paths from attributes of root node to classification result of leaf nodes, respectively define different rules for classification. Eventually, it builds a decision tree with tree structure to judge user's possible next activity.

2) Procedure

Suppose that x_i represents an attribute for an item. The attributes set of an item X is donated as $X = \{x_1, x_2, x_3, \dots, x_n\}$. There are numbers of classes, which are represented as $Y = \{y_1, y_2, \dots, y_m\}$.

- For each attribute x_i in current training set, respectively, calculate their rate of information gain (Gain Rate).
- Select attribute x_i with the largest Gain Rate as root node of current decision tree.
- For x_i , classify samples with the same value of x_i into one subset
- For each subset, if the current set includes positive and negative samples, then recursively call the algorithm.
- If the current set includes either positive or negative sample, then indicate the subset as leaf node, and mark corresponding branch as P or N. Return to caller.

3) Gain Rate Calculation

- Category Gain Rate

$$I_E(C) = - \sum_{j=1}^m P(C_j) \log_2 P(C_j)$$

$P(C_j) = \frac{|C_j|}{|C|}$ is frequency of sample $c = c_j$ in the whole training rate.

- Category Gain Rate under a specific attribute

$$I_E(C|X) = - \sum_{j=1}^n \sum_{i=1}^m P(x_i) P(C_j|x_i) \log_2 P(C_j|x_i)$$

$P(x_i)$ is rate of number of sample $x = x_i$ to all samples, $P(C_j|x_i)$ is used for calculating frequency of sample $c = c_j$ under $x = x_i$ in the whole training set.

- Information Gain

$$Gain(X) = I_E(C) - I_E(C|X)$$

- Attribute Information Gain

$$I_E(X) = - \sum_{i=1}^n P(x_i) \log_2 P(x_i)$$

$P(x_i)$ is rate of number of sample $x = x_i$ in the whole samples.

- Gain Rate

$$gain_ratio = Gain(X)/I_E(X)$$

In C4.5, each node stores the information used for calculating Entropy, which is obtained by counting number of positive and negative samples for each value of attributes. According to information in nodes, attribute with minimum value of Entropy can be deduced to classify the object.

C4.5 algorithm has the following advantages [29][30]: Generated classifying rules are easier to understand; A higher accuracy rate exists. Its disadvantages are: During construction of the tree, data set is repeatedly scanned and sorted, which causes algorithm inefficiently. Additionally, C4.5 is limited on the scale of data. Data set has to reside in memory. If data training set is too large to be accommodated in the memory, then the program cannot run.

2. Naïve Bayes

This algorithm must first establish a model that describes the previous data set and concept set. The model is built by analyzing samples described by attributes (Or instance, objects, etc.). Assume that each sample has a pre-defined class which is identified by attributes and known as Class Label. Training set is composed by data elements which are used for analyzing model. This step is called as supervised learning.

Assume that set $X = \{x_1, x_2, x_3, \dots, x_n\}$, each x_i is a attribute for this item X . There are a variety of classes within the set $Y = \{y_1, y_2, \dots, y_m\}$.

In the well-known Bayesian formula (Equation 2.1), assume that A and B are two events, and $P(A) > 0$,

$$P(B|A) = \frac{P(AB)}{P(A)} \quad (2.1)$$

Using the Multiplication formula $P(ABC) = P(C|AB)P(B|A)P(A)$, Bayesian Formula can be transformed into

$$P(y_i|X) = \frac{P(Xy_i)}{P(X)} = \frac{P(X|y_i)P(y_i)}{P(X)} \quad (2.2)$$

Set of x_i is denoted by X , which is called as the attribute set. General relation between X and Y is uncertain. So we can only say how much degree X might belong to y_i , like X belongs to y_i by 80%. Then X and Y are regarded as random variable, $P(Y|X)$ is called posterior probability of Y . Relatively, $P(Y)$ is prior

probability of Y .

In the training phase, according to data collected from training data, for each combination of X and Y , the posterior probability of $P(Y|X)$ is trained. When classifying, for a sample of X , it discovers the largest y_i in a bunch of posterior probability $P(Y|X)$ which is obtained from former training and determines which class X belongs to. The denominator $P(X)$ is ignored because it is constant. Prior probability $P(Y)$ is estimated by proportion of each sample in the whole training set. In order to calculate $P(X|Y_i)$, X 's each attribute $x_1, x_2, x_3, \dots, x_n$ should be independent, then the formula

$$P(X|Y_i) = \prod_i P(x_i|Y_i)$$

can be used.

In the Bayes algorithm, for the same Y , the largest $P(y_i|X)$ is extracted to determine the class of X .

2.3.2 Algorithm dependent on Confidence

The above two algorithms have good prediction result on the student platform. However, the correctness is still wished to be improved. So Confidence is introduced to measure the prediction result.

1. Confidence Calculation

By constructing models, some predictions appear. But how can we believe them? For example, a result illustrates that a student will attend the TOEFL exam under the condition of bachelor degree, intended country-USA and so on. So should it be convinced? Confidence is a measurement that provides a standard for making a decision.

The confidence value of each prediction result should be calculated. Re-model all the elements might affect confidence to acquire the confidence of the prediction value. For each given sample data, its class is deduced under these models. Then the prediction by corresponding confidence is described. 0 represents that the result is completely uncertainty; 1 represents that prediction is very credible. Confidence is a value ranged between 0 and 1. Confidence of a sample with predicted class will be produced under the Confidence Model.

1) Procedure of building Confidence

First, delete the correct class attribute which is "type" from the original data set, and add two new properties of "forecast" and "confidence" to the original data set. The algorithms are run to predict "forecast". Then "forecast" is filled with the prediction result. Finally, prediction values are compared with correct expected value. If the two is same, then "confidence" is 1, if not, then 0.

2) Procedure of training Confidence

New data set contains original common attributes, "forecast" and "confidence" is analyzed by an algorithm which can process numbers to form a new model – Confidence Model. The common attributes and "forecast" are prediction conditions and "Confidence" is the result should be predicted. This model can be used for

extrapolating “confidence” by inputting personal information and “forecast”.

2. Improved with Confidence

Confidence indicates whether the result can be trusted. Thus, each sample with pre-condition should produce the classified result under Prediction Model. Then the sample with the classified result is processed under Confidence Model to acquire its confidence. As for the same classified result, the one of higher confidence is regards as one which is more close to direction of user behavior. Therefore, the prediction with higher confidence is the final result for usage. In the next chapter, an experiment is conducted to prove its effectiveness.

2.3.3 Standard for Comparing Algorithm

Measures of Classifier Evaluation [30][32] are the following:

1) TP Rate

$$TP_{Rate} = \frac{P}{P_{ALL}}$$

P is number of correctly classified positive samples, P_{ALL} is number of all the samples which were divided into positive samples

2) FP Rate

$$FP_{Rate} = \frac{N_P}{N_{ALL}}$$

N_P is number of negative samples which are misclassified(i.e., classified as positive samples), N_{ALL} is number of all the samples which were divided into negative samples.

3) Precision

$$Precision = \frac{P}{P + P_N}$$

P is number of correctly classified positive samples, P_N is number of positive samples which are misclassified (i.e., classified as negative samples).

4) Recall

$$Recall = \frac{P}{P + N_P}$$

P is number of correctly classified positive samples, N_P is number of negative samples which are misclassified (i.e., classified as positive samples).

5) F-Measure

$$F = \frac{2 * Precision * Recall}{Precision + Recall}$$

This measure is related to last two.

6) Correct instance

$$C = \frac{P}{I_{All}}$$

P is number of correctly classified positive samples, I_{All} is number of all the samples.

Above measures except FP Rate analyze the classifier's rate of positive samples. FP Rate studies the validity of model in the aspect of wrong division.

2.4 Hypertext Modeling

Hypertext Model [31] represents web application's modules and organizational structure. The model mainly consists of two parts: Composition Model and Navigation Model. Composition Model depicts Page and its Content Unit. Page is the carrier in which application provides users with information. Unit is the basic atomic content element in the data model. Navigation represents link between units in Page.

Among them, content unit contains data unit and index unit. Data unit contains two basic elements. Source entity specifies entity property contained in data unit. Data unit represents one entity object while entity may have multiple objects. Index unit represents more than one object instances of entity in a list. In this thesis, hypertext model is applied to show link between Pages, providing the basis for implementing prototype system.

2.5 Summary

This chapter elaborated an analysis approach based on process and machine learning. Process comes from typical experience of human being doing a transaction. In a process, each phase represents a function module in some ways. Therefore, this approach applied the vertical and horizontal process model to analyze users' different phases and their detail activities, and extract concept of user objects, services and phase classification. According to detail activities obtained from process analysis, related data and its properties were studied, in order to provide the basis for machine learning. Machine learning is to study how computer simulates or implements human being's behavior, in order to acquire new knowledge or skills and reorganize the existing knowledge structure to keep improving their performance. In this chapter, the Decision Tree and Naïve Bayes were discussed. The hypertext model was introduced to provide the basis of prototype system.

3. User Behavior Analysis based on Student Platform

Last chapter presented the proposed analysis method based on process and machine learning. In this chapter, target users will be analyzed in four aspects, namely, process modeling, data attribute analysis, algorithm modeling and hypertext modeling.

3.1 Process Model

In this thesis, the student platform supports overseas students who are divided into three categories: high school students, undergraduate, graduate, according to their different educational background. Among them, undergraduate and graduate play the most important roles of potential overseas students in China. Of course, there are some young students applying for middle school and high school. However, this part of people apply colleges by receiving services from some other persons or agencies instead of on their own, so they are not target group in this thesis. Among the three types of service targets, people apply for graduate school abroad will be the largest group of people adopting web application services. Because they need to collect their own data, collate information and submit an application.

3.1.1 Concept Analysis

This approach is a process-based behavior analysis. Process is a kind of human being's inertia expression. People behave in a certain order after they repeat for a long time. A single process is not enough for expressing human's behavior. Process can be classified as short-term and long-term one. In this thesis, short-term process represents some atomic and tightly executed activities. Long-term process is combination of phases which are in a chronological order. And each phase is composed by activities.

After analyzing features of target people, it is found that students want to study abroad will experience four phases, i.e. Language Examination, College Application, Visa Application and Private Affairs Preparation (Table 3.1).

First of all, the phase of Language Examination is checked. Background material contains a serial of material about exam and requirement of applied country, which is aimed at informing students with a basic understanding of requirement of language exam, such as, language certificate, degree-of-difficulty factors and so on. Exam application and attendance are tasks in this phase, which are necessary for every student who wants to study abroad. By contrast, Tutorial class Application and Tutorial occur as candidate tasks in the phase of Examination, which are not compulsory. So is Hotel Booking.

In the phase of College Application, students will focus on the background material for colleges, majors, mentors and intended countries instead of exam which had been finished. Meanwhile, in order to assist students to apply appropriate

colleges, a large number of skill materials about application are demanded.

In the phase of Visa Application, students require information about consulate and visa application, and prepare their own material for certificate.

Table 3.1 Phases and Activity

Phase	Activity
Language Examination	Background Material Exam Type Decision Exam Application Tutorial class Application Tutorial Hotel Reservation
College Application	Background Material Application Help Material Express
Visa Application	Consulate Material Application Help Material Express
Private Affairs Preparation	Flight Ticket Booking Private Affairs Arrangement and Purchase Help Material Car Rent

In the last phase of Private Affairs Preparation, students give prominence to Flight Ticket Booking, Private Affairs Arrangement and Purchase, some aboard guidance and vehicle arrangement in the day of departure. Among them, Ticket Booking requires extracting a reasonable recommendation from a large number of flights, which are the most troublesome problems for students. As for private goods, a list of necessities aboard should be proposed as well.

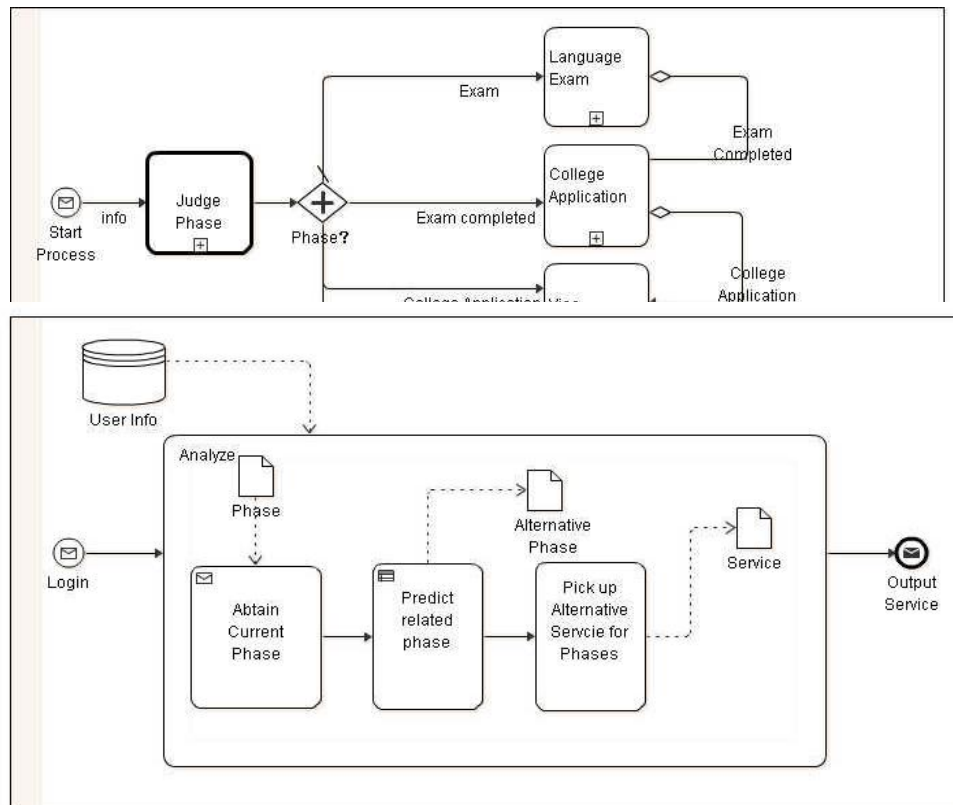
3.1.2 Horizontal process modeling

This thesis takes student platform as a precedent and proposes a vertical and horizontal mixed process model. Horizontal process represents different phases of applying college. By the preceding analysis, it is found that students should experience four main phases of Language Examination, College Application, Visa Application and Private Affairs Preparation. These phases are in a certain order. It says that the former phase is the pre-condition of the following phase from the perspective of a student.

Figure 3.1 illustrates the model of horizontal process for the platform. In the beginning, it goes into a module of phase judgment (Figure 3.2). Language Examination is the default phase for the process. If the phase is failed to judge, the

system will be start from the Language Examination. If other result appears, then jump to that one. Every phase can be executed as a beginning phase. The current one is transformed into the next one until it completes. That means, for students, there is prerequisite relation among four important phases.

Figure 3.1 Horizontal Process Model



Therefore, according to relation of time or logic, the candidate of certain phase is its later phase. The previous phase is irreversible for the current one. For example, when a student passes the language exam, he goes into the phase of College Application. Exam score is the premise of his application for the college. According to the general rule, a student in the phase of language exam is most concerned about information of exam. Then exam result may affect the direction of students' application. So the phase of college application is a candidate associated phase.

Figure 3.2 Phase Analysis

Figure 3.2 depicts the detailed process of Phase Analysis. First, the system collects users' current status. In the prototype system, they might receive a short question about their status. Users' basic information can be obtained from database. The module of Prediction related Phase studies the phase related to the current one, which is its following phase in the horizontal process. Then it extracts activities (services) of the phase from database. A collection of services are returned as output.

3.1.3 Vertical Process Model

In the proposed horizontal and vertical process model, the vertical process represents users' detailed activities in each phase, which will be transformed into prototype's functions from the model. The Process Diagrams of four phases are displayed as bellowed.

As Figure 3.3 says, Language Examination phase contains a background survey before exam. According to the student's self-condition (economy, school, major, score) and intended country, the exam goals (exam type, score level) are determined. After registering exam, some suitable tutorial classes, tutorial (material on the Internet) and hotels for exam are recommended by considering type of the selected exam, language and so on.

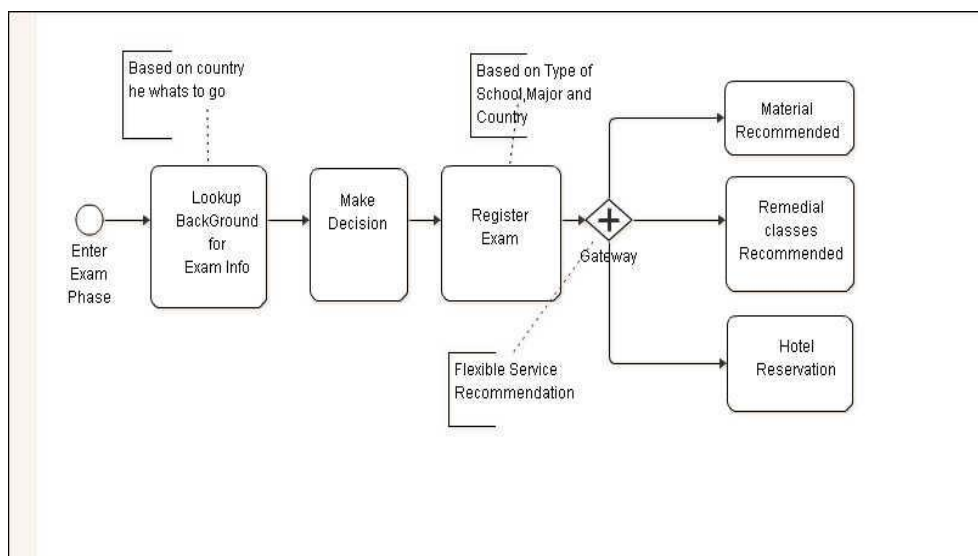


Figure 3.3 Language Exam Process Diagram

College Application Phase (Figure 3.4) contains a survey of intended country and college (Major, College, and Mentor). Basic services are recommended by the student's self-condition (economy, college, major and score) and intended country. Then the student goes to stages of Applying, and result waiting. In the last stage, if receive Offer from college, the student determines whether takes in. if failed, the process will go back to the beginning.

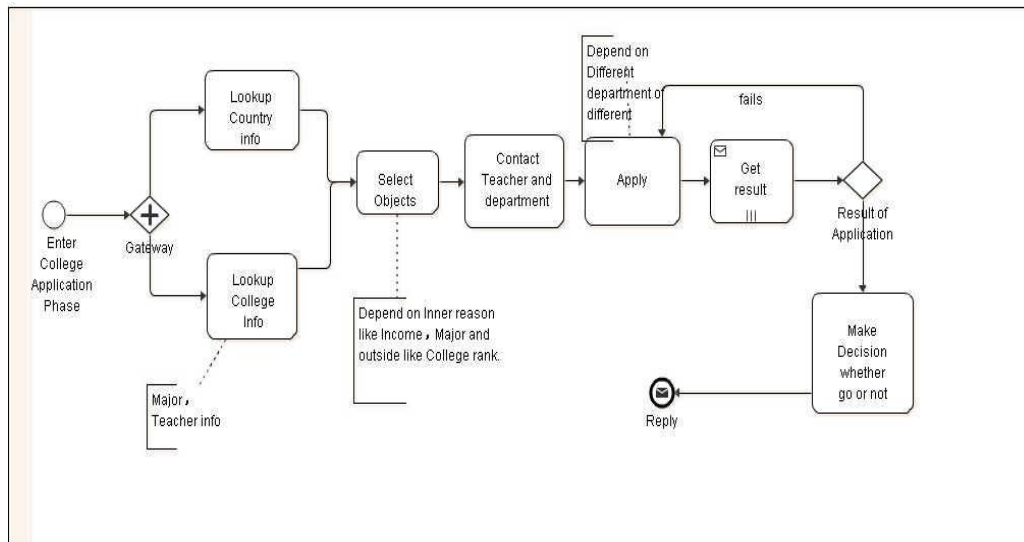


Figure 3.4 College Application Process Diagram

Visa Application Phase (Figure 3.5) contains a survey of application material, dependent on the success of college application. Material is recommended by requirements on visa for the country in which the college locates. During the preparation of material, four sub-processes are executed in parallel. There is no requirement on sequence. It is the same as the waiting link in the College Application Process. If the student applies successfully, the process completes, otherwise she/he might reenter the activity of material preparation link.

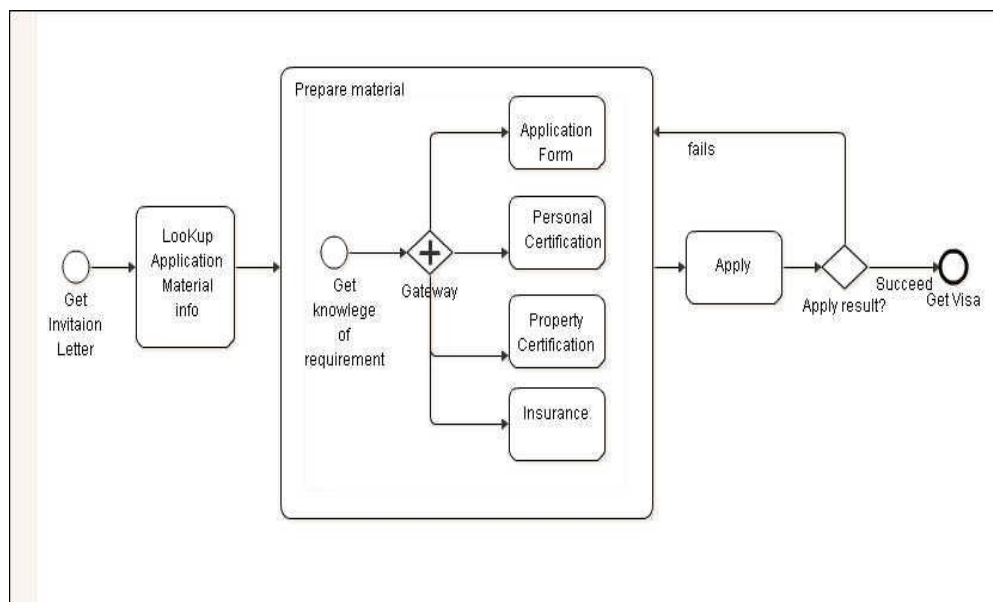


Figure 3.5 Visa Application Process Diagram

The activities of last Private Affairs Preparation (Figure 3.6) are relatively spare. In this phase, services are recommended based on the Preparation List. The services are divided into Booking Fight Ticket, Buying Affairs and Changing Money. The prediction results in this phase are processed based on user's choice of all the former

phases, such as the country which the student will go.

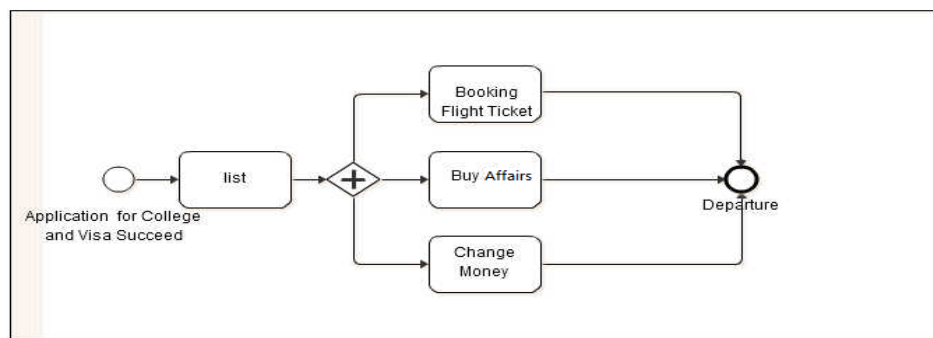


Figure 3.6 Private AffairsPreparation Process Diagram

3.1.4 Collaborative Process Model

Collaborative Process Model is used for representing the information interchange and collaboration relationship between two processes.

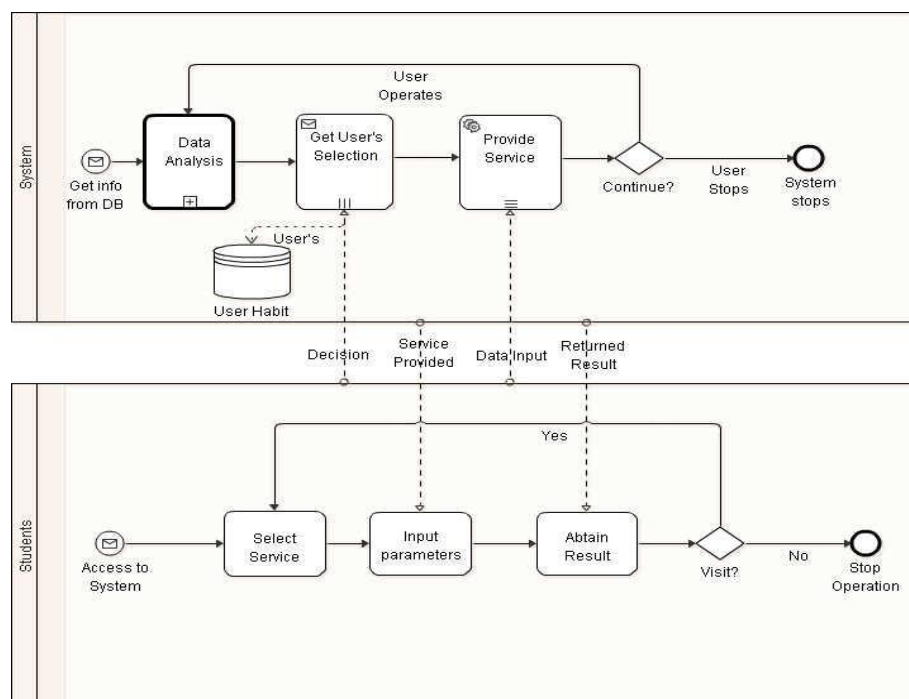


Figure 3.7 Collaborative Process Diagram

In Figure 3.7, list the interchange relationship between System and Student and their behavior process model. Seen from Lane Students, for a student, his/her behaviors involved in using websites, consist of selecting service, inputting information and obtaining result which are the same as activities of the process of visiting ordinary website. However, in the Lane System, there is one more activity of Data Analysis, providing services during the system's behavior process, besides receiving input and selection from users. Actually in the prototype system, some of user's behaviors are captured and recorded for further prediction (in Section 4.4)

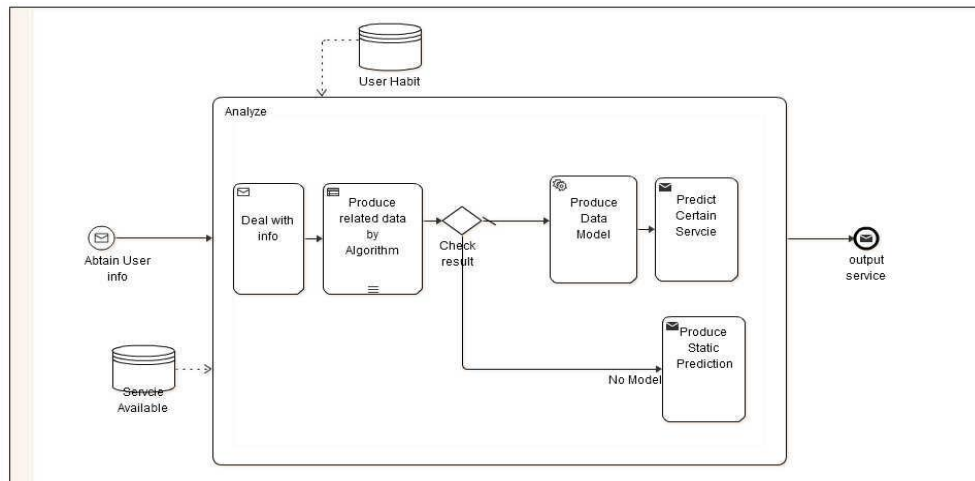


Figure 3.8 Data Analysis Sub-process Diagram

Figure 3.8 describes the structure of Data Analysis which is a sub-process for the Collaborative Process. In this process, the system will pick up User Habit from Database and deal with user-related information by algorithms of machine learning in order to produce a data model and predict user's services might need. In case that the data model is unable to construct owing to lacking of data and an unsuitable algorithm, the system will return some collections of pre-setting services.

3.2 Data Model

Data model is also constructed in four aspects of Language Private Affairs Preparation, College Application, Visa Application and Private Affairs Preparation. In this thesis, data model pays attention on the relationship between two entities. Entity property influences the generation of instance and provides the basis for Prediction Model.

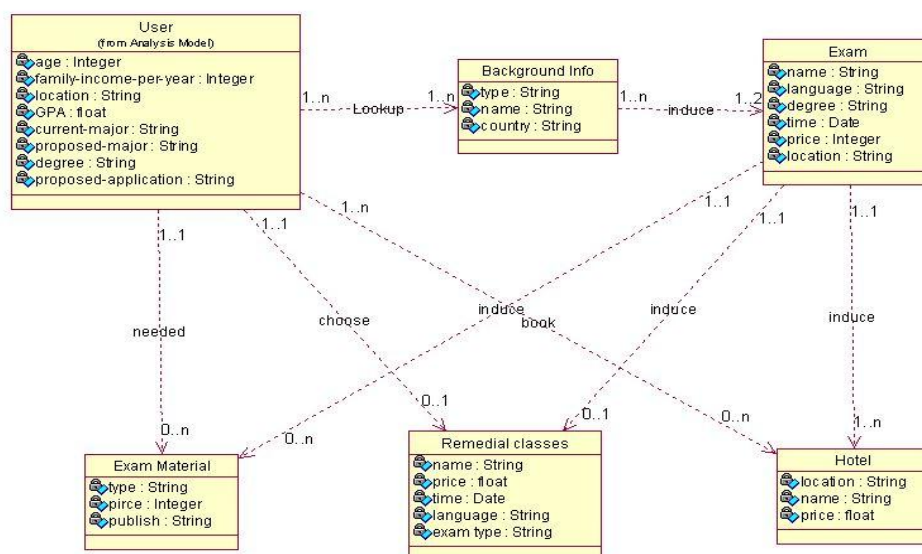


Figure 3.9 Language Exam Data Model

As for the language exam phase, figure 3.9 mainly displays the related data to this phase. Background info embodies three properties: Type, Name, and Country. Type includes comprehensive information of country (economic power, foreign student amount, working environment), language exam information (difficulty, score, required by which exam), these properties are in reasoning relation with Exam. Check Exam's properties, language and name are connected with the instance of Exam Material, language, name and time are linked with the instance of Remedial Classes, location and time can infer some recommended hotel. Meanwhile, some properties of a global user instance like family-income-per-year determines relation among three instances.

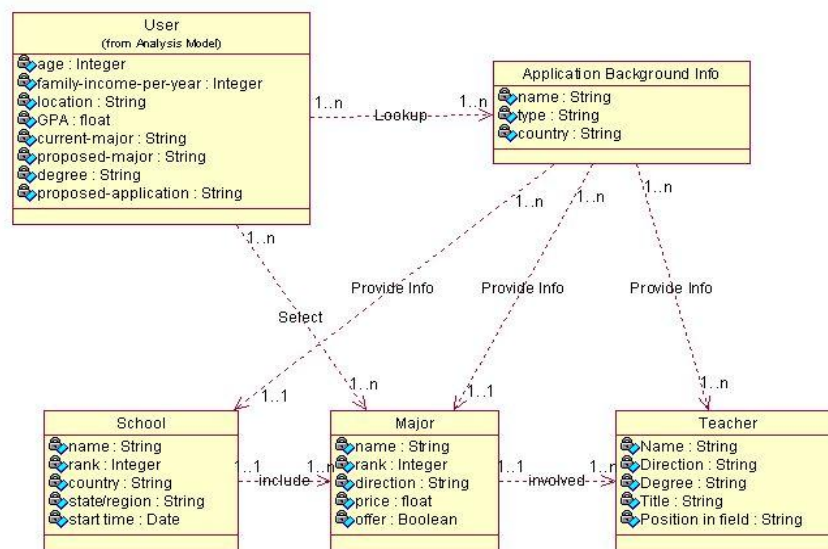


Figure 3.10 College Application Data Model

In College Application phase (Figure 3.10), a user determines his/her major according to Application Background Info. School keeps 1 to N with Major while Major is 1 to N with Teacher as well. School, Teacher and Major are the conjunct factors affects user's decision.

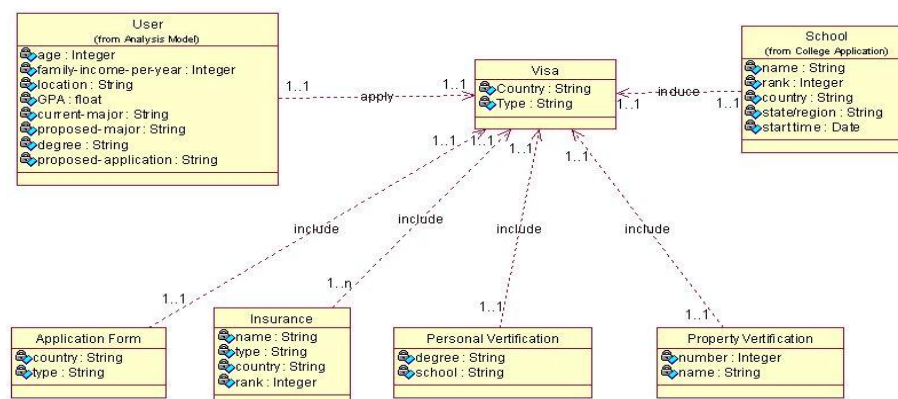


Figure 3.11 Visa Application Data Model

In the Figure 3.11 and 3.12, it finds that related entity mainly points at Visa and Private Staff. Visa has three entities: Insurance, Personal Verification and Property Verification. During Visa Application, the precondition of Visa is related to Applied College, especially the country where the college locates, which is a key factor affects procedures of Visa Application.

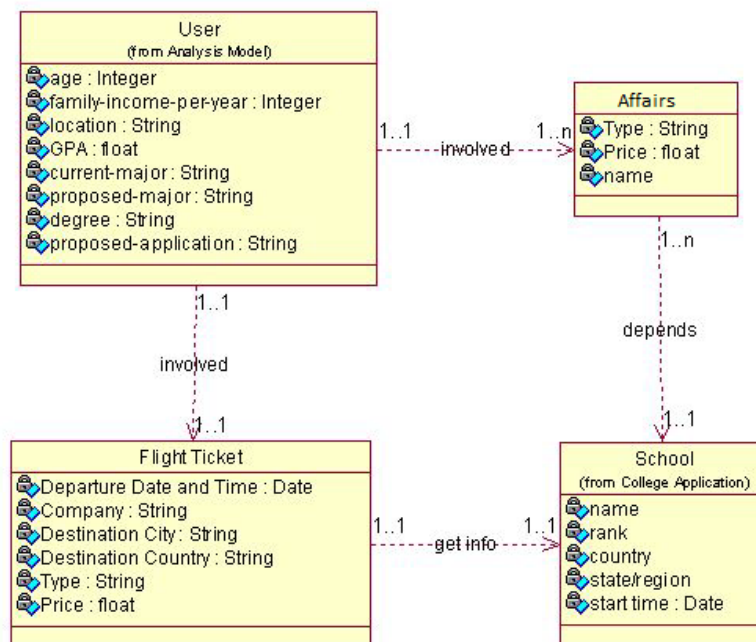


Figure 3.12 Private Affairs Preparation Data Model

In a system, almost all the data associating with each other, while data is related between phases which can be identified from above four phases. For example, the global user instance affects all the instances produced by data, which is decided by the basic rule of human activity. Human's subjective idea and some outside conditions determine following activities for each process. That's why required data is different. Even for the same person, under various conditions, his/her decision is different. For instance, it is found that the country with higher consumption is welcomed by those who have higher salary. Moreover, intended country is likely to be changed with exam type.

Data model in this section lists corresponding entity's property and relationship, and provides attributes for algorithms of machine learning.

3.3 Prediction Model Based on Machine Learning

Through continually accumulation and practice, human being forms a series of sequence activities in the area of studying and life. These activities in certain order represent some general rules in human behaviors. It finds that properties of each activity event associates with appearance of the next event. How to mine these rules is an important study topic. Machine learning as one of aspects in the data exploring, takes more attention on the implementation of algorithms to mine suitable data and rules. Various improved algorithm can capture human behavior in some aspects and

prediction. In this chapter, two algorithms for classification are studied: Decision Tree and Naïve Bayes.

In this thesis, a survey is conducted for users' basic information. As the limit of personal time and resource, its scope is in the phase of Language Examination. User's selections on exam are involved.

3.3.1 Data Resource

As one of main factors to affect the correctness of rules or behaviors, historical data plays an important role in machine learning. Thus, a survey for group of target users is first step of collecting user information.

1. Goal

This thesis mainly emphasizes students who want to study abroad, aimed at analyzing a series of behavior activities in the process of applying college, in order to provide some personal services and reduce their effort on retrieving information. After a thorough procedure of Process Model and data analysis, it is clear about the properties or instance what are required by Prediction Model. From that view, a questionnaire is made to collect information and serve experiment on algorithm.

2. Object

The survey object in this thesis is undergraduate, graduate and some students in high school. The first two groups of people occupy a large part of the whole group, which is easier to get answer.

3. Survey Approach

The original data resource in this thesis is users' data obtained from questionnaires. Questionnaires are sent out by email, or posts on social websites, or communication groups.

Mainly people receive this questionnaire are selected by following rules that

1) They should have experience of studying abroad, or 2) They are in the process of applying, or 3) They have intention of studying abroad.

Some key questions in a questionnaire are

- 1) Age
- 2) Sex
- 3) Degree(Higher)
- 4) Major
- 5) Degree(Want to apply)
- 6) Income per year (Family)

In the survey, most people touched upon are my friends, classmates, other known people and someone connected by forum or other tutorial class. As the limit of personal effort, I invite five people to be involved in this survey and help sending out questionnaires.

Because parts of questionnaires are sent electronically, lot of questionnaires cannot be recycled, such as those questionnaire sent by email. Some of interviewees

cannot assure their intended country, and then their answers are useless for later analysis. Thus after collating the recycled data, there are 258 effective answers.

4. Data Analysis format

Data will be input into original data information center after filtering and analysis. All the data will be transformed into ARFF format in order to train and construct the prediction model by Weka.

The data format of Weka is ARFF (Attribute-Relation File Format), this is a kind of ASCII text file. The whole ARFF file can be divided into two parts. The first part lists Head Information, including the announcement of relationship and attributes. The second part lists Data Information, which is in the data set.

Attributes announcement represents as the beginning of “@attribute”. For each attribute in the data set is corresponding to certain “@attribute”, which defines its name and data type.

```
@relation weather

@attribute outlook {sunny, overcast, rainy}
@attribute temperature real
@attribute humidity real
@attribute windy {TRUE, FALSE}
@attribute play {yes, no}

@data
sunny,85,85,FALSE,no
sunny,80,90,TRUE,no
overcast,83,86,FALSE,yes
rainy,70,96,FALSE,yes
rainy,68,80,FALSE,yes
rainy,65,70,TRUE,no
overcast,64,65,TRUE,yes
sunny,72,95,FALSE,no
sunny,69,70,FALSE,yes
```

Figure 3.12 Data set's Arff format example

Data information is behind the flag of “@data”. “@data” occupies one line. From its next line is information of each instance which occupies one line individually. Each attribute is divided by “,”. “?” represents attribute which is missing value and cannot be ignored. Weka supports four types of data: numeric, nominal, string, date.

3.3.2 Analysis tool

In this thesis, Weka[33] is applied for analyzing and predicting which is an open-source data mining work platform and integrate a large amount of machine learning algorithm taking on tasks of data mining, including data pre-processing, classification, regression, clustering and association rules. It provides a visual interface for user to operate on.

Figure 3.13 shows the Interface after opening a file of training set which contains its statistics result, data properties, and result type.

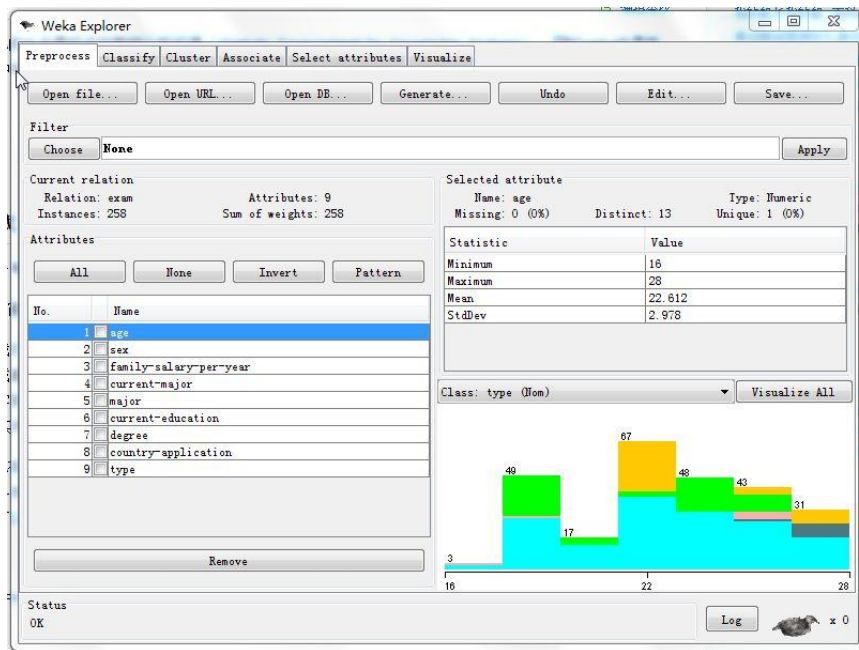


Figure 3.13 Weka Main Interface

Figure 3.14 gives the Interface of Weka's classification. Users are allowed to select certain classification algorithm. In this thesis, apply J48 of Weka for Decision Tree C4.5. At the right side of interface, list some performance measures for algorithm: Correct instance, FP Rate, TP Rate, Precision, F-measure which are used for comparing these classification algorithms.

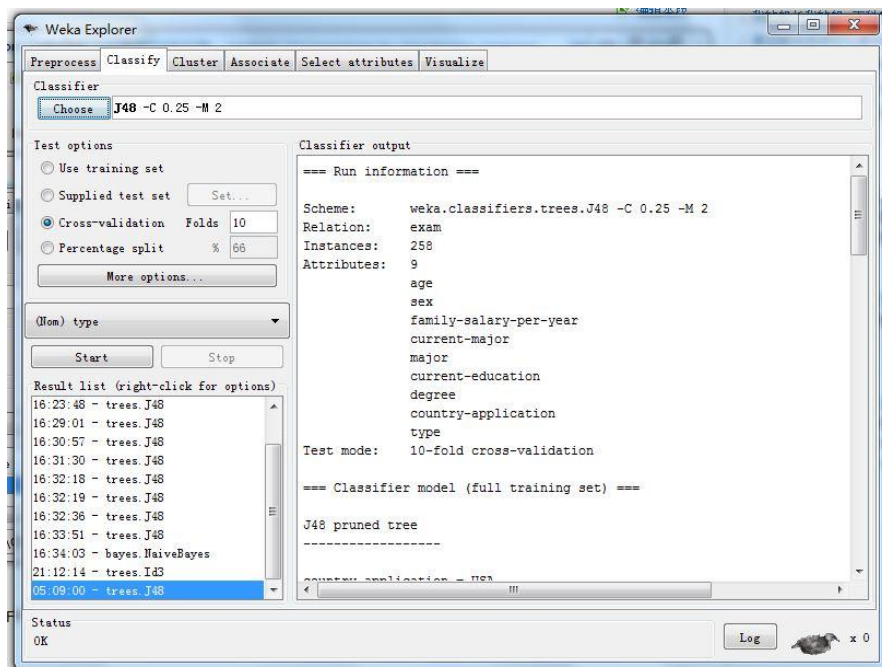


Figure 3.14 Weka interface for classification

Besides visual interface, Weka provides corresponding APIs which can be extended by programming. Each algorithm is represented as a class such as

LinearRegression, DecisionTable, J48, NaiveBayes and so on. They are inherited from AbstractClassifier which implements classifyInstance and predicted according to completed model.

1. Main interface

Classifier, contains buildClassifier(Instances data) which models data set of Instance by selected classifier, and classifyInstance(Instance instance) predicts unknown type of instance according to the built model.

Instance as interface of managing instance, declaims all the operation related to data set. Such as attribute(int index), Attribute classAttribute(), double classValue(), Instances dataset(), int numValues(), int numAttributes() and so on.

2. Main Class

Evaluation contains all classes related to model verification and algorithm.

Instances represents data Structure of dataset in the memory, including operation on the dataset such as adding, modification, deleting and sorting.

```

Evaluation evaluation = new Evaluation(dataSource);
classifier.buildClassifier(dataSource);
abclassifier = classifier;
evaluation.crossValidateModel(classifier, dataSource, 10,
    new Random(1));

System.out.println(classifier.toString());
System.out.println(evaluation
    .toSummaryString("\nResults\n\n", true));

```

Figure 3.15 Sample Code

Predict Logic of Prototype system in this thesis is implemented based on related APIs on Weka.

3.3.3 Data Collection and Analysis

Information of the students who study aboard or have intention of that is collected by questionnaires. Data is filtered and collated in order to make prediction more reliable. The result is transformed as a training set in Arff format which is required by Weka to model and predict under Decision Tree and Naïve Bayesian

1. Basic Attributes for students

Some attributes related to prediction for an ordinary student might have age, sex and family-salary-per-year which are basic background features. In some classical data set, those attributes will be collected. For example there are a dataset for predicting whether an American will consume more than \$50 per month. Other attributes related to personal study are GPA, Current-Major, Major (intended), Degree (Current), country-application.

Table 3.2 Student Personal Attributes

No	Name
1	Age
2	Sex
3	Family-salary-per-year
4	GPA
5	Current –major
6	Major
7	Current-education
8	Degree
9	Location
10	Country-application

In personal information, country-application will determine where final schools which students apply are. Personal current major and intention major determine the type of exam and school-application. Assume that individual's address or region, to some extent, determines individual's some decisions. There exists the traditional view that Beijingers like New York, Shanghai people like Tokyo and so on. Attributes in Table 3.2 work for questionnaires design and produce an input set of classification algorithm.

2. Category Set for Examination Phase

It finds, in the collected data, that main exams students attend include IELTS, TOEFL, GRE, GMAT, SAT. As the special nature of the abroad examination, parts of students require multiple exam score. Therefore, the possible candidates can be classified into:

Table 3.3 Statistics on Result Category

No	CLASS	COUNT	WEIGHT
1	TOEFL	8	8
2	IELTS	152	152
3	SAT	6	6
4	GRE	0	0
5	GMAT	0	0
6	TOEFL+GRE	55	55
7	TOEFL+GRE+GMAT	37	37

Because of the generality and function of exam is diverse, some exams' result is useless alone and they have to work with other course such as GRE and GMAT. SAT is an exam which provides a language quiz for those senior high school want to apply college. As the limit of the interviewee, we find that number of SAT is 6.

3. Data Attributes Processing

In the experiment, data will be analyzed as Table 3.4. The first 8 attributes come from personal attributes in Table 3.2 will be prediction conditions. The **type** is the prediction result of exam's type.

Table 3.4 Attributes for Examination Phase

No	Name
1	Age
2	Sex
3	Family-salary-per-year
4	Current –major
5	Major
6	Current-education
7	Degree
8	Country-application
9	Type

Based on Table 3.5, the raw data from research is processed into a training set of 258 samples. The last column is each sample's final category result. This set will be the input of classification algorithm and model of predicting some sample case.

Table 3.5 Attribute Value

Attributes	Value
Age	Number
Sex	Female,Male
Income	<50K,50K-100K,100K-300K,>=300K
Current-major	no-major,engineering,science,arts,literature,business,medicine
Major	engineering,science,arts,literature,business,medicine
Current-education	high-school,bachelor,master,phd
Degree	high-school,bachelor,master,phd
Country-application	USA,UK,France,Germany,Spain,Italy,Switzerland,Sweden,Danmark, Netherlands,other-european-country,Canada,others

3.3.4 Prediction Model based on C4.5 and Naïve Bayes

This part will give pictures of the models produced by two algorithms. These two algorithms try to figure out rules of condition attributes and conclude which category the certain sample belongs.

The training set got from analysis step of 3.4.3 are input into Weka and trained as two algorithms C4.5 and Naïve Bayes. 10 fold Cross-Validation [30] is applied. The data were randomly divided into 10 parts. 9 of them are used for building the prediction model. The model is listed as:

1. C4.5 Model

Apply C4.5 to deal with the data. The Decision Tree is produced as Figure 3.16. The root of tree is the **country-application** which obeys the normal rule of Examination. In the reality, students select exam after they determined which country they want. Thus, different exam is applied in different country. For example, USA requires TOEFL while general European country demands IELTS. In some cases, these two basic language exams are in common use. However, as the reason of the difference on score transforming, students seldom do like that.

```

country-application = USA
|  major = engineering
|  |  sex = Female: SAT (5.0/1.0)
|  |  sex = Male
|  |  |  age <= 25: GRE+TOEFL (18.0)
|  |  |  age > 25: GRE+TOEFL+GMAT (2.0)
|  major = science: GRE+TOEFL (19.0)
|  major = arts: GRE+TOEFL (0.0)
|  major = literature: GRE+TOEFL (0.0)
|  major = business: GRE+TOEFL+GMAT (36.0/1.0)
|  major = medicine: GRE+TOEFL (12.0/2.0)
country-application = UK: IELTS (22.0)
country-application = France: IELTS (21.0)
country-application = Germany: IELTS (23.0/1.0)
country-application = Spain: IELTS (6.0)
country-application = Italy: IELTS (8.0)
country-application = Switzerland
|  age <= 25: IELTS (19.0)
|  age > 25: TOEFL (2.0)
country-application = Sweden: IELTS (12.0/2.0)
country-application = Denmark: IELTS (19.0)
country-application = Netherlands
|  age <= 27: IELTS (13.0)
|  age > 27: TOEFL (2.0)
country-application = other-european-country: TOEFL (3.0/1.0)
country-application = Canada
|  sex = Female: IELTS (8.0)
|  sex = Male: GRE+TOEFL (7.0)
country-application = others: IELTS (1.0)

Number of Leaves :    23
Size of the tree :    30

```

Figure 3.16 Decision Tree for Exam

When **country-application** is USA, tree is divided as Major. The rule is followed the regulation of USA college which requires more than one language exam depended on different major. But some other countries like UK has no special requirement.

2. Naïve Bayes

According to formula 3.1 concludes the highest one with $P(Y_i|X)$. Y_i 's priori probability is in the Table 3.6. Distribution of X list in Appendix A. For each input case data, it will be speculated by the highest $P(Y_i|X)$.

Table 3.6 priori probability

Y_i	$P(Y_i)$
GMAT	0
GRE	0
IELTS	0.58
TOEFL	0.03

SAT	0.03
GRE+TOEFL	0.21
GRE+TOEFL+GMAT	0.14

3. Algorithm Validation and Contrast

This thesis applies 10 Fold Cross-Validation [30] to verify the above model. The data were randomly divided into 10 parts. After modeling and analyzing 9 of them, the last one is used for correctness validation on model. The evaluation is calculated by the average of 9 ones. The condition of the last data part is input. According the new model, it will get a result (Type). Compare this result with the Type recorded in the original dataset and check whether it is correct.

a	b	c	d	e	f	g	<-- classified as
0	0	0	0	0	0	0	a = GMAT
0	0	0	0	0	0	0	b = GRE
0	0	115	6	11	14	6	c = IELTS
0	0	2	6	0	0	0	d = TOEFL
0	0	0	0	6	0	0	e = SAT
0	0	1	0	0	52	2	f = GRE+TOEFL
0	0	1	0	0	0	36	g = GRE+TOEFL+GMAT

Naïve Bayes

a	b	c	d	e	f	g	<-- classified as
0	0	0	0	0	0	0	a = GMAT
0	0	0	0	0	0	0	b = GRE
0	0	149	1	1	1	0	c = IELTS
0	0	8	0	0	0	0	d = TOEFL
0	0	0	0	4	1	1	e = SAT
0	0	3	0	0	52	0	f = GRE+TOEFL
0	0	0	0	0	2	35	g = GRE+TOEFL+GMAT

C4.5 Tree

Figure 3.17 Contrast on Naïve Bayes and C4.5

Figure 3.17 is the comparison of two algorithms' classification results. Both of them have some errors. NB classified some results which should be IELTS as 6 TOEFL, 11 SAT, 14 GRE+TOEFL and 6 GRE+TOEFL+GMAT. Comparatively, C4.5 has obviously less errors.

Table 3.7 Statistics on C4.5 and Naïve Bayes

Measure	C4.5	NB
Correct Rate	93.0233 %	88.7597 %
TP Rate	0.93	0.888
FP Rate	0.066	0.022
Precision	0.905	0.928
Recall	0.93	0.888
F-measure	0.917	0.899

Seen from Table 3.7, it finds that correctness rate of NB is 88.7597% and C4.5 is 93.02%. According to TP Rate and Recall, judgment on positive samples of C4.5 is better than NB. But from the view of FP rate, C4.5's rate of misjudgment on negative samples is slightly higher than NB. When check Precision, it finds that C4.5 classified more negative samples as positive ones than NB did. However, as for the amount of

error and correctness rate and F-measure, C4.5 is better than NB for this training set.

NB algorithm takes advantage of mathematical theory. There is a solid mathematical foundation and stable classification efficiency. The algorithm requires few parameters and is less sensitive to missing data and. But NB has a premise that attributes of thing should be independent which is hard in the reality. Moreover, another key point of this algorithm is natural distribution of thing attributes. It is difficult for us to set it through assumption of data set's distribution probability. Once number of attributes and relation between them reach a certain extent, the algorithm's efficiency will be reduced. In addition, NB is based on mathematical statistics. In this thesis, because data collection is limited and might be not enough for NB.

In contrast, rule of the behavior that student selecting exam is simple. One of the advantages of C4.5 is quick setup of data analysis model based on small amount of data. There is fewer pre-process on user's input. Heuristic criteria of information entropy protect the optimal rate of information gain on each branch. Generated rule is simple and easy to understand. As the sample in this thesis, C4.5 is enough for keep the correctness of rule.

Both algorithms have their own merits in this situation.

3.3.5 Improved Prediction Model based on Comparison of confidence

1. Improved Method

In the section 3.4.2, after comparing model of C4.5 and NB, conclude that C4.5 and NB still have some improved points. According to TP Rate, FP Rate, Precision, Recall and F-measure, both two algorithms have advantages on judgment of positive and negative samples. We hope that the correctness of predict can be higher.

Thus, this thesis proposes an improved model based on contrast of confidence. Depending on completed prediction model of C4.5 and NB, compare confidences of their predictions. Select one with the higher confidence as prediction result. The detail algorithm is in the section 2.4.2.

2. Experiment Step

- 1) First, build two basic prediction models by C4.5 and NB under the testing dataset. Then obtain the Type (classified result) for samples in the validation dataset respectively. Meanwhile acquire correctness rates of two models.
- 2) Construct a new testing dataset by adding Confidence to each sample. Here confidence comes from the rule in section 2.4.1. Then build two Confidence Models with the new dataset for C4.5 and NB by MSP [27]. Calculate the confidence of validation dataset though two Confidence Models. Finally, compare the confidences of two classified result for each sample, and select the higher one to produce a new prediction (Type).
- 3) Compare the new Type with original one in the original data set and get the new correctness rate.

In this thesis, 90%, 70%, 50% of original training set are adopted to produce

models and 100% data are to validate. Detail algorithm is checked in Appendix B.

3. Experiment Result

Seen from Table 3.8, compared in vertical view, the correctness of model is increased by size of samples. In the horizontal view, improved model is higher than other two in the correctness rate.

Table 3.8 Contrast on Improved Model and Original

Method \ Model	C4.5	Naïve Bayes	Improved Model
90%	0.961240	0.914729	0.976744
70%	0.945736	0.918605	0.965116
50%	0.920635	0.857143	0.972868

Therefore, this thesis will adopt the improved model for analyzing user behavior in the prototype system in the next chapter.

3.4 Hypertext Model

Hypertext Model is the foundation of designing the web page logic. It displays the navigation of the new prototype system.

Figure 3.18-3.21 depicts the relevant page and component of each phase. Before entering the personal main page, the user has to provide the current state of applying. System will jump to certain phase's page. User's personal page includes certain service under his current phase and related phase. Section 3.2.2 illustrates the idea of related phase. After receiving the user's operation on his/her personal main page, the response page for his/her choice will be processed by Data Process and Prediction Model. That means the system filters services before publishing them to the user.

In Figure 3.18, the user determines which service he/she wants first. Suppose he/she selected Exam. Then, the prediction model checks the most possible type of exam he/she would attend and offer relevant service. Page goes to Info Page. The Country info is background information of user's intended country, such as employment, economic condition, education level or other exam information like difficulty, the country, school, degree required by this exam. After the user made decision on exam, he/she visits Exam Page, registers exam through certain service interfaces. Complete the exam operation, the system will recommend tutorial material and tutorial class as the formal operation information.

In Figure 3.19, the user can select required information like college and country or jump to corresponding detail introduction page. Data analysis module will predict suitable information. In Figure 3.20, the user faces with recommended page for Visa. There are several procedures or materials required by applying visa. Now in the system will list Verification and Insurance information

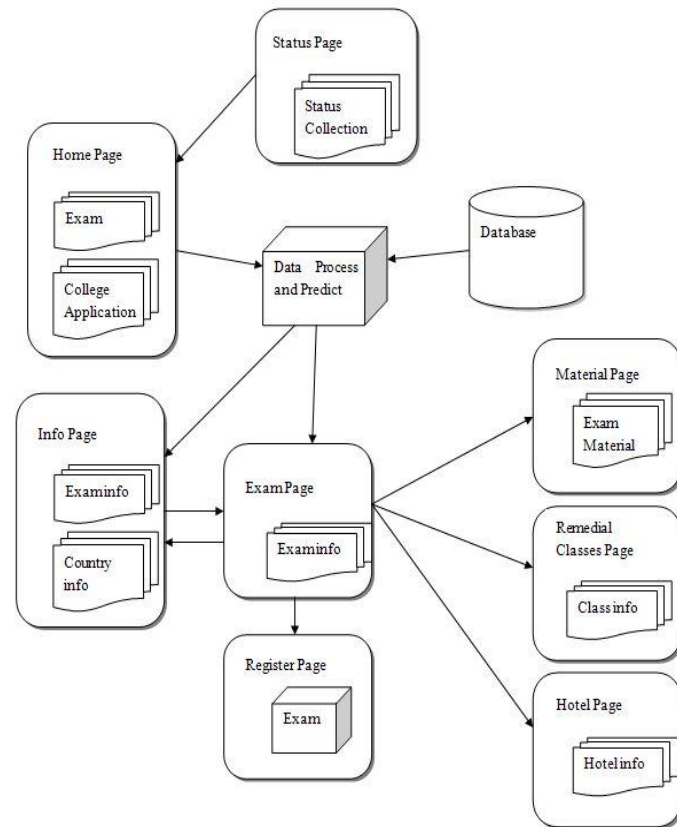


Figure 3.18 Navigation of Exam

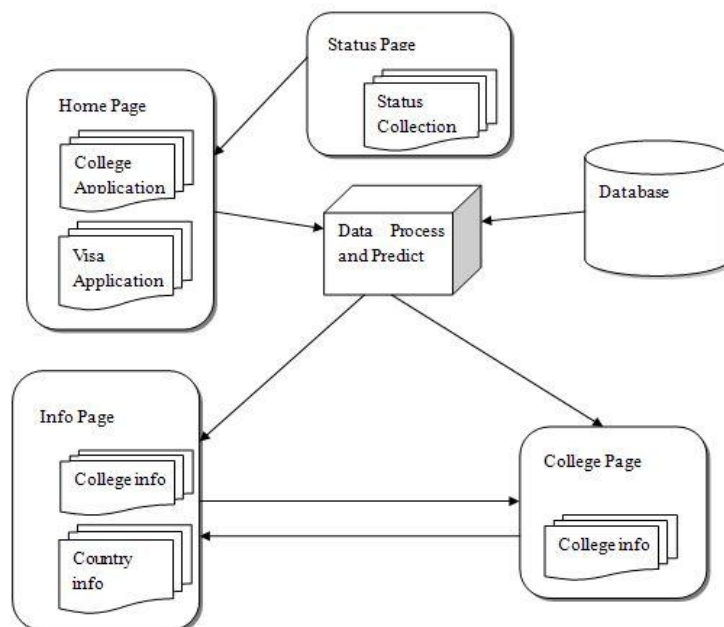


Figure 3.19 Navigation of Applying Phase

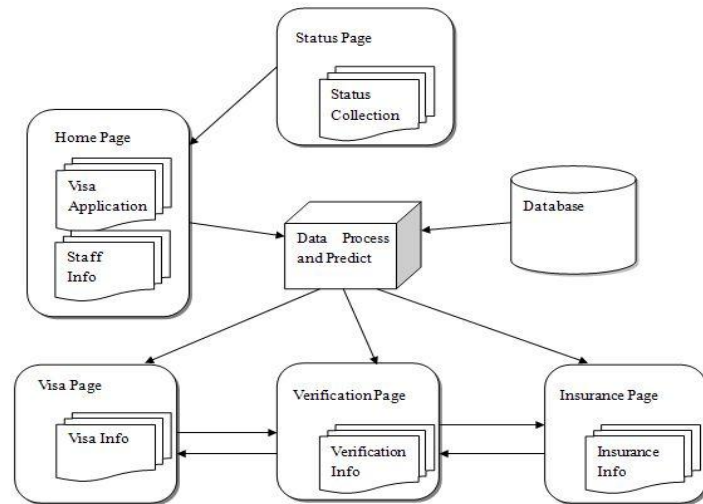


Figure 3.20 Navigation of Visa Phase

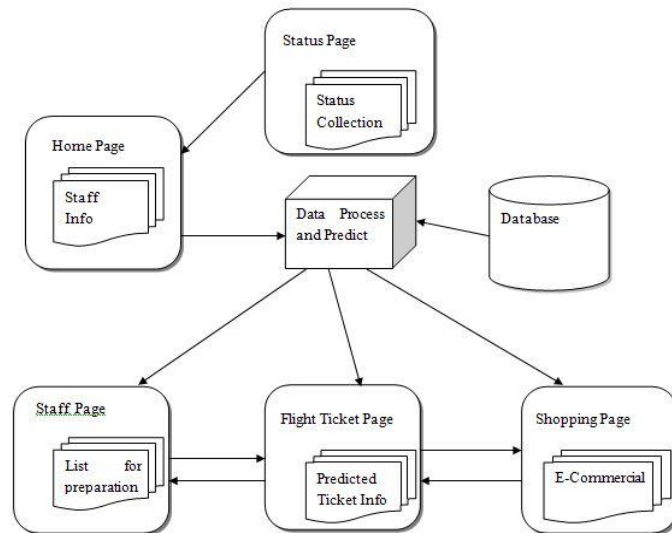


Figure 3.21 Navigation of Private Affairs Preparation

In Figure 3.21, in the phase of Private Affairs Preparation, the user obtains basic list of things for going aboard, which includes a flight ticket recommended with analysis of country, depart time, depart place, target location, price, and a shopping page with recommended affairs collected from certain portal websites,

3.5 Summary

This chapter applies the approach of behavior analysis proposed in last chapter, on student platform for those who want to study aboard, to build horizontal and vertical process model. Then, the model produces detailed phases and services. Finally, the collaboration process model between the system and users is built.

The process is analyzed to conclude main data objects and relation between them. The data collected from the survey is reorganized to sum up attributes. C4.5

and Naïve Bayes is conducted to predict users' behaviors. They are evaluated under five measures: TP Rate, FP Rate, Precision, Recall, F-Measure. An experiment is on improved model carried out to valid it.

In the last part, hypertext models are constructed on the basis of former analysis. Its main function is to abstract the architecture of the whole prototype system, and display the main component and link-relation between each page.

4. Prototype System based on User Behavior Analysis

In this chapter, a prototype system based on the approach proposed in the last two chapters will be introduced. It is web application providing services for students who will study abroad. Students input their personal information as the system requires, then it will recommend services for different people in different phases according to their clicking or other operations on the page. The system is to validate the user behavior analysis method proposed in the thesis.

4.1 Architecture

This prototype system is implemented for verifying whether the recommendations given by the User Behavior Analysis can satisfy users' requirements. Thus it owns four main function modules: Language Exam, College Application, Visa Application, and Private Affairs Preparation. Those modules are obtained from the analysis on users' process model (Section 3.2).

Figure 4.1 illustrates the main page after information process. The login user's intended country is USA and Major is business. The system concentrated on Language Exam to conduct the prediction for users' next behavior.

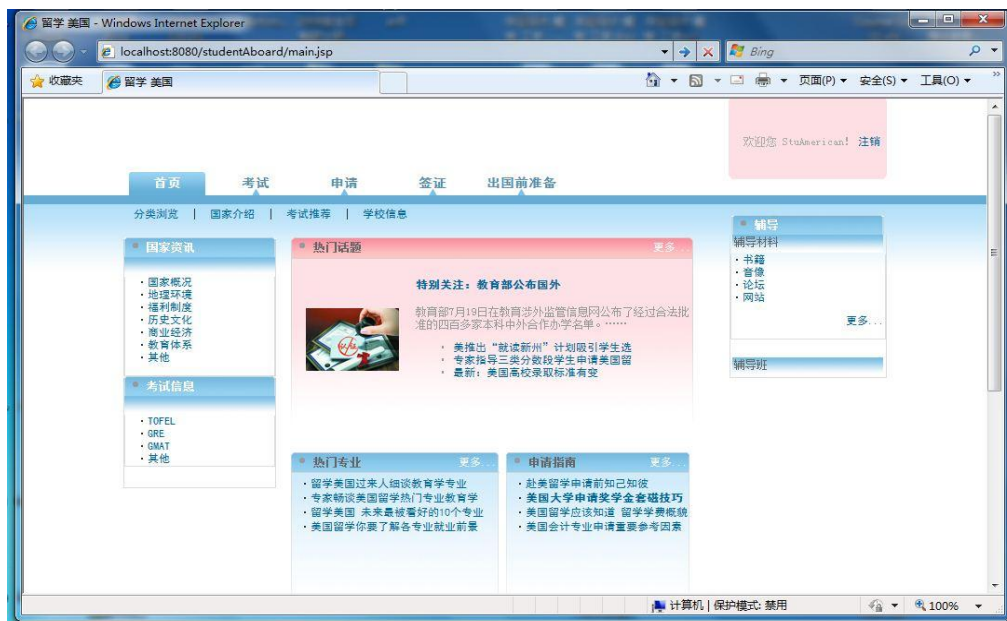


Figure 4.1 MainPage

4.1.1 System Module

MVC is applied in the system to deal with data, interface and logic control. The system is divided into 3 system modules: Core System, Decision System, and Statistics System.

The Core System provides UI page to serve students' different requirement in

the phase of Language Exam, College Application, Visa Application, and Private Affairs Preparation. It implements some basic functions like information publishing,

Besides normal component like Database, the key part is Decision System which contains the Prediction model. Decision system is in charge of speculating the service which might be required by the current user. Statistic System contains the basic data for building the prediction model. DB stores the accumulation of the data collected in the process of usage. Thus all the service users obtain are determined by the Decision System and Statistics System.

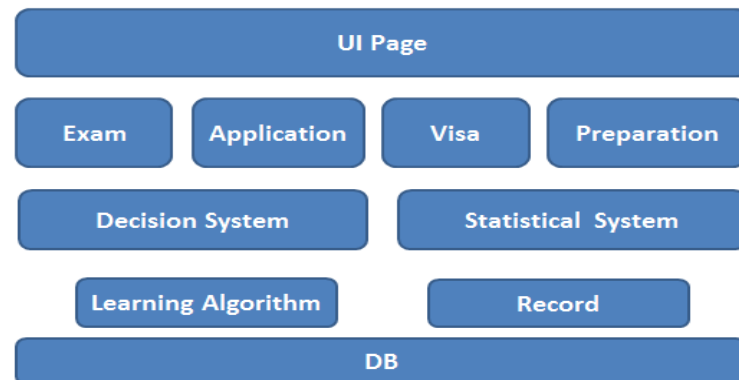


Figure 4.2 Architecture of Prototype System

1. Decision System

Decision System analyzes recorded data by certain machine learning algorithm, produces a prediction model and speculates a service which is suitable for users. The core algorithm for Decision System is improved based on Confidence. Its details are described in Section 2.4 and 3.4. The original model is calculated from training set given by the Survey. Every night, it will be updated with the new user data collected in the daytime.

2. Statistics System

Statistics System collects users' operations and habit. The detail will be in Section 4.4.

3. Database

DB records all kinds of training set and user's personal usage experience such as which kind of recommendation is popular, which recommended content or service is confirmed by users. The information is used for further investigating user behavior to acquire user's interests and habits.

System's decision module applies the prediction model produced by confidence analysis, acquires user's basic information from DB, and determines the service should be available in this phase. Figure 4.1 is the result by processing users' location and intended country.

4.1.2 Network topology

Considering the requirement of calculation capability and system resource,

modules should be distributed to three servers

The Core System and Statistics System are located in the Server. Decision System and Database are separated into two servers. Thus, Server can deal with a large number of visitors. The high calculation of Model Building in Decision System will not influence responses from Server. Database stores data including users' behavior history.

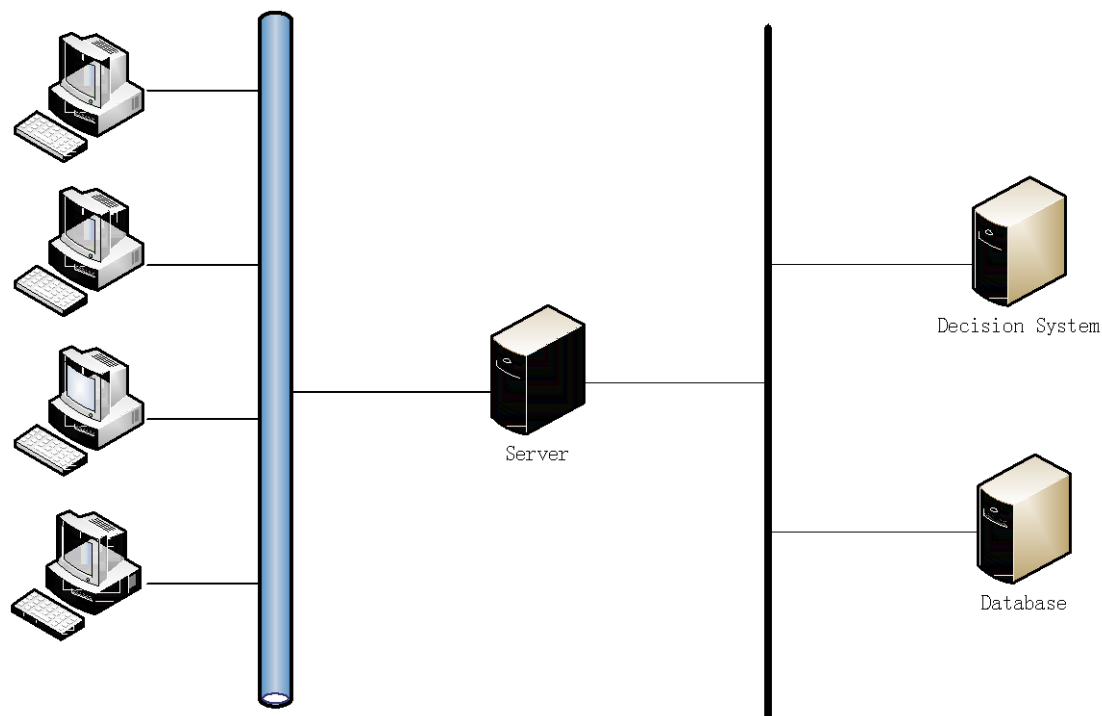


Figure 4.3 Network Topology

However, in the experiment, there is not enough resource to support the topology. Moreover, the efficiency of system is not main point for us to check. Thus, three servers are simplified as one.

4.2 System hardware and software environment

The hardware and software environment of the system in the experiment as:

- Intel i3 processor, 4GB memory.
- Windows 7 professional.
- Dreamweaver 8.
- Apache Tomcat.
- MySQL.

4.3 User Behavior Capture Module

Some users' operations [34] on the web are browsing web page, searching information, finding and saving article, clicking certain buttons and so on. In order to capture user's operation, there are two methods for tracing [35][36]:

1. Data from Server side

This method is achieved by reading log files in the server side (The second line in Figure 4.3). When a user visits the website, the server will record his/her operation, ID, login time and so on. Relatively speaking, the server is most likely to log some response records for request from client side. By adjusting the setting of web server, the server can record more kinds of required log information. Helpful information is processed and stored into database which can be available for later usage.

The disadvantage of this method is that there are some difficulties to extract information from log file. The log file contains various records like operation or exception from database which might be necessary for error check. Meanwhile, summing up user behavior from log file is an indirect analysis method. Thus, some behavior like button clicking cannot be captured.

2. Data from Client side

Client side is the original place where user behavior is produced, in which the system can capture user's browsing path and time more accurately. Nowadays a large amount of third parties provide service to analyze user behavior by embedding plug-in into client side. The plug-in does some statistics on flow of web and user's clicking. Statistics software is installed in the server side for monitoring the website. It is a convenient way of getting detail statistics information. Moreover, it costs only traffic of visiting statistics software.

Another method is to monitor by software installed by third party website. But it always costs a lot. Some websites provide service free. However, there will be limit in function. Thus in this thesis, this method is not suitable.

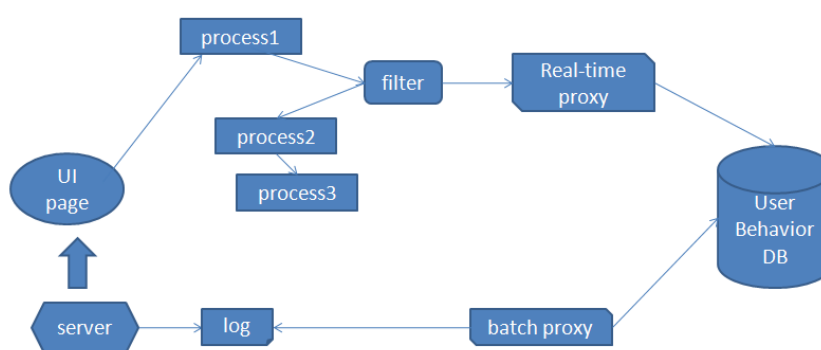


Figure 4.3 User Behavior Tracing Logic Diagram

3. Implementation

This thesis collects user's information in three ways: Register, Status Update Capture Business Logic.

1) Register

User's register information includes sex, age, current-major, intended major,

intended country and which phase user belongs to. Above information provides some basic elements for further analysis and predictions on user's behavior which can be seen from the process definition and model construction. That is because prediction model regards the basic personal information as prediction condition.

2) Status Update

The user will experience a short survey page for his/her current status, which is helpful for locating the entering phase of main page.

3) Web Behavior Capturing

Besides some static information like above, some users' dynamic operation on web should be captured and recorded as part of prediction conditions and historical data for updating model.

This thesis carries out the approach (Upper line in Figure 4.3) by adding business logic in coding to trace user operation on pages. The system gets user's choice on pages and calls relevant business operation logic from Servlet in the background. In the processing period of Servlet, it deals with the statistics information of users by Filter in Figure 4.3.

Compared to the method of analyzing log file in the server, this approach is detecting in real-time. Thus it burdens the system heavier. However, it is more accurate and efficient. Considering that this system is an experimental platform for validating, the burden of communication can be ignored.

4.4 Experiment

15 testers who have experience of applying college are invited to use the system. Here make comparison on system recommendations and testers' hope.

No.	age	sex	family-salary-per-year	current-major	major	current-education	degree	country-application	type
	Nominal	Nominal	Nominal	Nominal	Nominal	Nominal	Nominal	Nominal	Nominal
1	25.0	Male	50K-100K	arts	literature	bachelor	master	France	IELTS
2	23.0	Female	50K-100K	business	business	bachelor	master	Netherlands	IELTS
3	19.0	Female	<50K	no-major	engineering	high-school	bach...	Sweden	TOEFL
4	22.0	Female	50K-100K	medicine	business	master	master	UK	IELTS
5	21.0	Male	100K-300K	science	medicine	bachelor	phd	USA	GRE+TOEFL
6	24.0	Male	50K-100K	engineering	medicine	bachelor	phd	Germany	IELTS
7	24.0	Male	50K-100K	engineering	engineering	master	phd	USA	GRE+TOEFL
8	23.0	Female	50K-100K	arts	arts	bachelor	master	France	IELTS
9	23.0	Female	<50K	business	literature	bachelor	master	UK	IELTS
10	18.0	Male	<50K	medicine	medicine	bachelor	master	USA	GRE+TOEFL
11	18.0	Male	50K-100K	science	engineering	bachelor	master	USA	GRE+TOEFL
12	22.0	Male	50K-100K	business	business	bachelor	master	UK	IELTS
13	22.0	Male	50K-100K	business	business	bachelor	master	Switzerland	IELTS
14	26.0	Male	<50K	literature	business	master	phd	Sweden	IELTS
15	18.0	Female	<50K	no-major	business	high-school	bach...	USA	SAT

Figure 4.4 User Information

4.4.1 Original User Data

The system acquires users' operation by user behavior capture module and recommends by Decision System. Figure 4.4 illustrates the 15 testers' information before they used the system, which is the reference for comparing with recommendations.

The attributes except “**type**” are input into system as register information. When testers request “Exam” page, the system will recommend related materials for the “type” of exam. The “type” is the reference for comparing with the result system recommends. In the prototype system, the original prediction model is built on the 258 samples from the survey in the chapter 3.

4.4.2 Experiment Result and Analysis

In order to list advantages of improved algorithm, 15 testers’ basic information is processed by the C4.5 and Naïve Bayes as well, besides the result given by the prototype system.

Check the Table 4.1, the red ones are wrong predictions. Two attributes “Prediction” and “Confidence” in Table 4.1 which represents the predicted value and confidence of the result.

Compared with original user data, there are two predicting errors for C4.5. One of them is about high school student. It is found that in the original training data set, something related to them is less than 10%. So no enough training sample might influence the result. Meanwhile, as for another error, in the original training set, there are 12 samples with result “TOEFL”. However, the tester’s expected choice is “IELTS”. Therefore, the new data is ambiguous to the old model.

Table 4.1 Result for Three Prediction Model

No	C4.5		Naïve Bayes		Improved	
	Prediction	Confidence	Prediction	Confidence	Prediction	Confidence
1	IELTS	0.995092	IELTS	0.993271	IELTS	0.995092
2	IELTS	0.858988	IELTS	1.015946	IELTS	1.015946
3	IELTS	0.981658	SAT	0.086571	IELTS	0.981658
4	IELTS	1.022278	IELTS	0.992314	IELTS	1.022278
5	GRE+TOEFL	1.060438	GRE+TOEFL	0.924068	GRE+TOEFL	1.060438
6	IELTS	0.940869	IELTS	0.994586	IELTS	0.994586
7	GRE+TOEFL	0.998473	GRE+TOEFL	0.986709	GRE+TOEFL	0.994586
8	IELTS	1.000133	IELTS	0.995901	IELTS	1.000133
9	IELTS	1.000133	IELTS	0.990999	IELTS	1.000133
10	GRE+TOEFL	0.889476	GRE+TOEFL	0.985459	GRE+TOEFL	0.985459
11	GRE+TOEFL	1.00763	GRE+TOEFL	0.985459	GRE+TOEFL	1.00763
12	IELTS	1.002653	IELTS	0.992314	IELTS	1.002653
13	IELTS	0.88002	IELTS	1.017261	IELTS	1.017261
14	IELTS	0.744788	IELTS	0.987054	IELTS	0.987054
15	GRE+TOEFL+GMAT	0.918371	SAT	0.922827	SAT	0.922827

In this case, the result of NB is quite good. The only one error is with lower confidence “0.086571” which means unconvincing. Other correct ones are with relatively high confidence.

The result of “Improved” is given by the prototype system. According to Table 4.1, both C4.5 and NB obtain high rate of correct prediction. In the 15 results, the 15th sample’s NB confidence is 0.92 higher than C4.5’s 0.91. Its final “Prediction” for improved model adopts the value of NB. Then it is consistent with the original data. So the improved model keeps the high correctness rate. It finds that 14 of 15 samples are correct. The correctness satisfies most of the 15 testers.

Thus, it can be concluded that the method of behavior analysis in this thesis is valid for the students want go aboard.

4.5 Summary

In this chapter, a prototype system is designed and implemented which adopts the user behavior analysis method proposed in the thesis. The Decision System applies prediction model depending on improvement of Confidence analysis. 15 testers with background of studying aboard or applying college are invited. 14 of them acquire correct prediction. And the result’s confidence is all over 0.95 which means the result is convinced. Thus, it demonstrates that the model is valid.

5. Conclusion

5.1 Results

With the rapid development of web technology, web application's function is transformed from simple information publishing into interactive information sharing. Many varieties of information in the web are involved into people's life with the enlargement of the scale of web. Data mining and filtering become significant means of improving user experience on the web and promote the efficiency of information retrieval and recommendation.

Process, as a set of activities from human's long-term work, life and study, reflects some law of human behavior. The existing series of operational data from a user can extract some reasonable properties and produce some laws. Users find it easier to acquire services they want if information is filtered by above laws before published.

This thesis studies the current status of web in advance, combines with vertical and horizontal process modeling, takes advantage of machine learning algorithms to build a new user behavior analysis method.

The results presented in this thesis are:

1. Through analysis on trends of web application, some new requirements are figured out under interactive environment. Considering the defects of single process modeling proposed method with the Machine Learning.
2. According to the vertical and horizontal model, process is modeled in the respects of the abstract whole and the detailed part. Concluding the property relation between data for students studying aboard, and predicting the next behavior depending on historical data. Deliberating and comparing the Decision Tree and Naïve Bayes to report their differences. The improved method is proposed based on Confidence.
3. Implementation of a prototype system, and conducting an experiment for validation of the new method.

5.2 Future work

This thesis obtains some result based on the experiment and prototype system validation. However, there are still some work can be taken over in the future.

1. As the limit of time, the amount of data is not large. The prediction of student's choice of college, mentor and major will be interesting topic.
2. The target group in this thesis focuses on those who are in college. Those are only part of human going aboard, excluding worker, scholar and so on.
3. In this thesis, only two algorithms are applied. Actually there are hundreds of approaches existing. Other ones can be tested in the certain Scenario as well.

Acknowledgements

Graduate life passed quickly. In the past two years, I received so many help from different people.

I very appreciate my mentor Professor Liu Qin who gave me a lot of academic and technical guidance. I am so grateful to those who are involved in the Survey and send me lots of basic information.

I should give my continuous thanks to those who work together in the lab and those who join in the Sino-Swedish program. I cannot forget the moment we think and talk and difficulties we have conquered.

Many thank my reviewer Ivan Christoff in Uppsala University for his trust and any valuable suggestion on my report.

I own my heartily gratitude to my family and colleges in the company. All your supports make my study life go so well

References

- [1] HamidahJantan, AdbukHamdan, Zulaiha Othman, Human Talent Prediction in HRM using C4.5 Classification Algorithm, (IJCSSE) International Journal on Computer Science and Engineering, 2010.Vol 2(8):2526-2534.
- [2] KnaussE., LübkeD..Using the Friction between Business Processes and Use Cases in SOA Requirements.In Proceedings of REFS'08, 2008.
- [3] Smith H., FingarP.,Business Process Management. The Third Wave.Meghan-Kiffer Press, Tampa, FL, USA 2002.
- [4]VomBrocke J., Rosemann M.. Handbook on Business Process Management: Strategic Alignment, Governance, People and Culture (International Handbooks on Information Systems), 2010. Vol. 1.
- [5] Stephen A. Process Modeling Notations and Workflow Patterns. White of IBM Corporation,2006.
- [6] IanH.Witten, Eibe Frank. Data Mining Practical Machine Learning Tools and Techniques, 2005.
- [7] Hong Tzung-Pei, Tseng Shian-Shyong. Comparison of ID3 and its generalized version.Fourth International Conference on Computing and Information, 1992.
- [8] Ding Rongtao, JiXinhao, Zhu Linting, RenWei.Study of the Learning Model Based on Improved ID3 Algorithm. First International Workshop on Knowledge Discovery and Data Mining (WKDD 2008) , 2008: 391~395.
- [9] Zheng Yao, Peng Liu. R-C4.5 decision tree model and its applications to health care dataset. 2005 International Conference on Services Systems and Services Management, 2005.Vol(2):1099~1103.
- [10] Genc I., Diao R., Vittal , Decision Tree-Based Preventive and Corrective Control Applications for Dynamic Security Enhancement in Power Systems.Power Systems, IEEE Transactions on.2010 ,Vol 25 (3):1611 – 1619.
- [11] K. Bhaduri, R. Wolff, C. Giannella, and H. Kargupta.Distributed Decision Tree Induction in Peer-to-Peer Systems.Statistical Analysis and Data Mining, June 2008, Vol 1(2):85–103.
- [12] Lin, W., Alvarez, S., and Ruiz, C. Collaborative recommendation via adaptive association rule mining. In Proceedings of the International Workshop on Web Mining for E-Commerce (WEBKDD'2000).2000.
- [13]Will Hill, Larry Stead, Mark Rosenstein, George Furnas, Recommending and evaluating choices in a virtual community of use, Proceedings of the SIGCHI conference on Human factors in computing systems, 1995.
- [14] Li Chunlei, Study on recommendation of Electronic Bank Product based on Association Rules.[Master Thesis], 2010.
- [15] Meeyoung C.,HaewoonK.,Pablo R., et al. I tube, you tube, everybody tubes: analyzing the world's largest user generated content video system,Proceeding on IMC '07 Proceedings of the 7th ACM SIGCOMM conference on Internet measurement.2007.
- [16] Colette Rolland. Modeling the Requirements Engineering Process. 3rd European-Japanese Seminar on Information Modelling and Knowledge Bases. 1993
- [17] Chun Ouyang, Marlon Dumas.From business process models to process-oriented software systems.Journal ACM Transactions on Software Engineering and Methodology (TOSEM). 2009.Vol 19(1).
- [18] BRAMBILLA, M., CERI, S., et al. Process-modeling in Web applications.ACM Trans. Softw Eng. Methodol.2006.Vol 15(4): 360~409.
- [19] Gilles,O., Hugues,J..A MDE-Based Optimisation Process for Real-Time Systems. 13th IEEE International Symposium on Object/Component/Service-Oriented Real-Time Distributed Computing (ISORC), 2010.
- [20] Zhang Tian, Jouault,F.A MDE Based Approach for Bridging Formal Models.2nd IFIP/IEEE International Symposium on Theoretical Aspects of Software Engineering, 2008.

- [21]Murugesan,S and A.Ginige, A. Web Engineering: Introduction and Perspectives, Chapter 1 in "Web Engineering: Principles and Techniques", Idea Group Publishing, 2005.
- [22] Moreno N., Fraternalli.P.,Vallecillo A., A UML 2.0 profile for WebML Modeling, In ICWE '06: Workshop proceedings of the 6th international conference on Web Engineering, 2006:4.
- [23] Nora Koch, Alexander Knapp, Gefei Zhang and et al. UML-based Web Engineering: An Approach based on Standards . In Web Engineering: Modelling and Implementing Web Applications, 2008. 157-191
- [24]Roberto Acerbis, Aldo Bongio,.Web Applications Design and Development with WebML and WebRatio 5.0. Models and Patterns, 2008.
- [25] Tian Ai Kui.System Requirement Manage Model Sustain Web-Based Information Publishing System.International Conference on Apperceiving Computing and Intelligence Analysis, 2008:303~306.
- [26] Yu Jun, Hu Zhi-yi,.Requirements Modeling of Web-based Scheduling Information System. International Conference on Information Management, Innovation Management and Industrial Engineering, 2008.
- [27] J. Han and M. Kambert, Data Mining: Concepts and Techniques, Morgan Kaufmann, San Francisco, 2001.
- [28] Li hui, Hu xiaomei.Analysis and Comparison between ID3 algorithm and C4.5 algorithm in Decision Tree. Water Resource and Power.Vol 26(2): 129-132.
- [29] NiuXiaotai, JiaZhentang.Application on Negotiating about ID3 Reasoning Method.2008 International Symposium on Intelligent Information Technology Application Workshops, 2008:519~522.
- [30] Cao Wei, Zhang Naizhou. A C4.5 Decision Tree Based Algorithm forWeb Pages Categorization. Computer System Application .2010.Vol 19(10): 195-198.
- [31] FeiYui-Ku, Wang Zhi-jian.A concept model of Web components.Proceedings of 2004 IEEE International Conference on Services Computing, 2004.159~164..
- [32] AyseCufoglu, MahiLohi, KambizMadani.A Comparative Study of Selected Classifiers with Classification Accuracy in User Profiling.2009 WRI World Congress on Computer Science and Information Engineering. 2009.Vol 3:708~712.
- [33] Remco R. Bouckaert, Eibe Frank, Mark A. Hall, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten.WEKA-experiences with a java open-source project.Journal of Machine Learning Research, 2010.Vol 11:2533-2541.
- [34] Wang Xia.User Behavior Analysis System Based on Web Browse: [Master Thesis]. Beijing, 2010
- [35] Zhou Yue. User Behavior Analysis System Based on Interest Classification: [Master Thesis]. Beijing, 2010
- [36]Dong Fuqiang.Web User Behavior Analysis and Application: [Master Thesis].Xian.2005

Appendix A Distribution of Naïve Bayes

1. Class GMAT: Prior probability = 0

age: 0 Normal Kernels.

StandardDev = 0.1667 Precision = 1.0

Mean = 0

sex: Discrete Estimator. Counts = 1 1 (Total = 2)

family-salary-per-year: Discrete Estimator. Counts = 1 1 1 1 (Total = 4)

current-major: Discrete Estimator. Counts = 1 1 1 1 1 1 (Total = 7)

major: Discrete Estimator. Counts = 1 1 1 1 1 (Total = 6)

current-education: Discrete Estimator. Counts = 1 1 1 1 (Total = 4)

degree: Discrete Estimator. Counts = 1 1 1 1 (Total = 4)

country-application: Discrete Estimator. Counts = 1 1 1 1 1 1 1 1 1 1 1 (Total = 13)

2. Class GRE: Prior probability = 0

age: 0 Normal Kernels.

StandardDev = 0.1667 Precision = 1.0

Mean = 0

sex: Discrete Estimator. Counts = 1 1 (Total = 2)

family-salary-per-year: Discrete Estimator. Counts = 1 1 1 1 (Total = 4)

current-major: Discrete Estimator. Counts = 1 1 1 1 1 1 (Total = 7)

major: Discrete Estimator. Counts = 1 1 1 1 1 (Total = 6)

current-education: Discrete Estimator. Counts = 1 1 1 1 (Total = 4)

degree: Discrete Estimator. Counts = 1 1 1 1 (Total = 4)

country-application: Discrete Estimator. Counts = 1 1 1 1 1 1 1 1 1 1 1 (Total = 13)

3. Class IELTS: Prior probability = 0.58

age: 13 Normal Kernels.

StandardDev = 0.9733 Precision = 1.0

Means = 16.0 17.0 18.0 19.0 20.0 21.0 22.0 23.0 24.0 25.0 26.0 27.0 28.0

Weights = 1.0 1.0 20.0 7.0 1.0 12.0 38.0 20.0 10.0 24.0 1.0 2.0 15.0

sex: Discrete Estimator. Counts = 88 66 (Total = 154)

family-salary-per-year: Discrete Estimator. Counts = 86 58 4 8 (Total = 156)

current-major: Discrete Estimator. Counts = 16 36 28 12 22 35 10 (Total = 159)

major: Discrete Estimator. Counts = 48 10 9 15 59 17 (Total = 158)

current-education: Discrete Estimator. Counts = 16 76 63 1 (Total = 156)

degree: Discrete Estimator. Counts = 1 12 67 76 (Total = 156)

country-application: Discrete Estimator. Counts = 3 23 22 23 7 9 20 11 20 14 2 9 2 (Total = 165)

4. Class TOEFL: Prior probability = 0.03

age: 2 Normal Kernels.

StandardDev = 1.0607 Precision = 1.0

Means = 25.0 28.0

Weights = 1.0 7.0

sex: Discrete Estimator. Counts = 7 3 (Total = 10)

family-salary-per-year: Discrete Estimator. Counts = 2 8 1 1 (Total = 12)

current-major: Discrete Estimator. Counts = 1 1 1 1 9 1 1 (Total = 15)

major: Discrete Estimator. Counts = 1 1 1 4 6 1 (Total = 14)

current-education: Discrete Estimator. Counts = 1 1 9 1 (Total = 12)

degree: Discrete Estimator. Counts = 1 1 1 9 (Total = 12)

country-application: Discrete Estimator. Counts = 1 1 1 1 1 1 3 3 1 3 3 1 1 (Total = 21)

5. Class SAT: Prior probability = 0.03

age: 3 Normal Kernels.

StandardDev = 3.266 Precision = 1.0

Means = 17.0 18.0 25.0

Weights = 1.0 1.0 4.0

sex: Discrete Estimator. Counts = 7 1 (Total = 8)

family-salary-per-year: Discrete Estimator. Counts = 6 2 1 1 (Total = 10)

current-major: Discrete Estimator. Counts = 7 1 1 1 1 1 1 (Total = 13)

major: Discrete Estimator. Counts = 5 1 1 1 2 2 (Total = 12)

current-education: Discrete Estimator. Counts = 7 1 1 1 (Total = 10)

degree: Discrete Estimator. Counts = 1 7 1 1 (Total = 10)

country-application: Discrete Estimator. Counts = 7 1 1 1 1 1 1 1 1 1 1 1 1 (Total = 19)

6. Class GRE+TOEFL: Prior probability = 0.21

age: 9 Normal Kernels.

StandardDev = 1.0787 Precision = 1.0

Means = 18.0 19.0 20.0 21.0 22.0 23.0 24.0 25.0 26.0

Weights = 19.0 2.0 1.0 3.0 3.0 6.0 12.0 7.0 2.0

sex: Discrete Estimator. Counts = 14 43 (Total = 57)

family-salary-per-year: Discrete Estimator. Counts = 31 21 2 5 (Total = 59)

current-major: Discrete Estimator. Counts = 1 27 21 1 1 2 9 (Total = 62)

major: Discrete Estimator. Counts = 27 20 1 1 1 11 (Total = 61)

current-education: Discrete Estimator. Counts = 1 46 11 1 (Total = 59)

degree: Discrete Estimator. Counts = 1 1 35 22 (Total = 59)

country-application: Discrete Estimator. Counts = 48 1 1 2 1 1 1 1 1 1 1 8 1 (Total = 68)

7. Class GRE+TOEFL+GMAT: Prior probability = 0.14

age: 4 Normal Kernels.

StandardDev = 0.822 Precision = 1.0

Means = 22.0 25.0 26.0 27.0

Weights = 26.0 3.0 1.0 7.0

sex: Discrete Estimator. Counts = 8 31 (Total = 39)

family-salary-per-year: Discrete Estimator. Counts = 6 21 11 3 (Total = 41)

current-major: Discrete Estimator. Counts = 1 14 10 1 1 12 5 (Total = 44)

major: Discrete Estimator. Counts = 3 1 1 1 36 1 (Total = 43)

current-education: Discrete Estimator. Counts = 1 27 12 1 (Total = 41)

degree: Discrete Estimator. Counts = 1 1 19 20 (Total = 41)

country-application: Discrete Estimator. Counts = 38 1 1 1 1 1 1 1 1 1 1 1 (Total = 50)

Appendix B Implementation of Improved Algorithm

```
Instances mainSource, j48Source, nbSource;
String mainfile = "exam.arff";
String j48file = "exam-output-j48.arff";
String nbfile = "exam-output-nb.arff";

try {
    mainSource = DataSource.read(mainfile);
    j48Source = DataSource.read(j48file);
    nbSource = DataSource.read(nbfile);

    Instances result = nbSource;

    int mainNum = mainSource.numInstances();

    int attNum = j48Source.numAttributes();
    mainSource.setClassIndex(mainSource.numAttributes() - 1);
    j48Source.setClassIndex(attNum - 1);
    nbSource.setClassIndex(attNum - 1);
    result.setClassIndex(attNum - 1);

    int correctNum = 0;
    for (int i = 0; i < mainNum; ++i) {
        Instance j48Ins = j48Source.get(i);
        Instance nbIns = nbSource.get(i);

        Instance temp = j48Ins;
        if (j48Ins.classValue() >= nbIns.classValue())
            temp = j48Ins;
        else
            temp = nbIns;

        result.get(i).setValue(attNum-2, temp.value(attNum-2));
        result.get(i).setValue(attNum-1, temp.value(attNum-1));

        if(temp.value(attNum-2) == mainSource.get(i).classValue())
            correctNum ++;
    }
    DataSink.write("CompareResult.arff",result);
    System.out.println("The correctly rate is " + correctNum*1.0/mainNum);
} catch (Exception e) {
    e.printStackTrace();
}
```